TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Antti Häkkinen

Quantifying Transcriptional Dynamics and Their Effects on Genetic Motifs from Live Cell Fluorescence Microscopy

Julkaisu 1370 • Publication 1370

Tampere 2016

Tampereen teknillinen yliopisto. Julkaisu 1370
Tampere University of Technology. Publication 1370

Antti Häkkinen

Quantifying Transcriptional Dynamics and Their Effects on Genetic Motifs from Live Cell Fluorescence Microscopy

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB223, at Tampere University of Technology, on the 12th of February 2016, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2016

Supervisor: Associate Professor Andre S. Ribeiro
Tampere University of Technology
Finland

Pre-examiners: Assistant Professor Matthew R. Bennett
Rice University
United States of America

Assistant Professor Matthew Scott
University of Waterloo
Canada

Opponent: Dr. James C. W. Locke
University of Cambridge
United Kingdom

ISBN 978-952-15-3685-4 (printed)
ISBN 978-952-15-3702-8 (PDF)
ISSN 1459-2045

Abstract

Advances in measurement techniques based on fluorescent tagging have enabled visualizing individual transcripts and proteins over time as they are produced in live cells. Such methods are critical in understanding how genes and genetic networks function, how they respond to external signals, such as stress conditions and temperature changes, and how cellular aging and diseases can affect their performance. This is relevant, as the functioning of genes and genetic networks affects the survival of cells and cell populations.

However, as cellular processes are complex and inherently stochastic, statistical signal processing methods are required to pre-process, analyze, and interpret the results from the measurement data. This stems from various traits of the data: averages poorly describe the behavior of multimodal populations; confidence estimates and hypothesis testing must be used to compare results, as they feature significant variability; and large data sets are required such that comparison can be made with sufficient confidence, rendering manual quantification excessively laborious. In addition, when combined with stochastic models, such methods can be used to extract information, which cannot be directly measured with current techniques. Meanwhile, the methods must be carefully designed, in order to avoid hidden assumptions which obstruct the objective quantification and comparison of the results. Finally, the methods should be made robust against errors characteristic to the measurement system or propagating from the earlier stages of the analysis.

Here, methods were developed in order to enhance the amount and the quality of the information which can be extracted from single-molecule measurements of live cells. In particular, methods for estimating RNA numbers and RNA production intervals from static images of cell populations and from time series of images of growing cells were first established. Next, methods for estimating the subprocesses of transcription were developed, as these processes cannot be directly measured in live cells. Computer simulations and live single-RNA measurements were used to demonstrate the reliability and performance of the new methods, indicating that the methods can adapt to different measurement settings and can be applied to other similar dynamical estimation problems. Finally, computer simulations of genetic networks are used to demonstrate that the accuracy of such methods is

paramount, as, in the dynamical ranges extracted from measurement, changes in gene expression dynamics have implications on the behavior of genetic networks, which are reflected on the behavior of individual cells and of cell populations as a whole.

The outcomes of this thesis respond to the demand of carefully designed statistical methods for accurate and unbiased quantification and comparison of cellular processes. Advances in such methods are necessary in order to generate new insight on the dynamics and the regulatory mechanisms of gene expression from single-molecule, single-cell measurements in live cells. The methods and the findings presented here will be critical for the success of such studies, contributing toward understanding how changes in gene expression patterns influence the cellular aging, stress, and diseases.

Preface

This study was carried out at the Department of Signal Processing, Tampere University of Technology under the supervision of Associate Professor Andre Ribeiro.

Foremost, I would like to express my sincere gratitude to my supervisor, Andre Ribeiro. I am deeply grateful for his persistent guidance and support throughout my doctoral studies.

I would also like to thank all the numerous colleagues and collaborators I have had pleasure to work with. In particular, I would like to thank the persons who have helped this thesis to materialize: Huy Tran and Stefania Garasto for contributing in performing simulations and in the theoretical aspects of the works; Meenakshisundaram Kandhavelu, Anantha-Barathi Muthukrishnan, and Jarno Mäkelä for providing me with measurement data for the publications and for experimenting; and Jason Lloyd-Price for discussions which have influenced some of the design features of the methods.

Finally, I would like to thank Tampere City Science Foundation, Jenny and Antti Wihuri Foundation, and Alfred Kordelin Foundation for financial support during my studies.

Tampere, August 2015,
Antti Häkkinen

Contents

List of abbreviations

CDF	Cumulative distribution function
GFP	Green fluorescent protein
CLT	Central limit theorem
CME	Chemical master equation
CRLB	Cramér-Rao lower bound
DNA	Deoxyribonucleic acid
EM	Expectation maximization
LR	Likelihood ratio
MC	Monte Carlo
ML	Maximum likelihood
ODE	Ordinary differential equation
PDF	Probability density function
PRNG	Pseudorandom number generator
RFP	Red fluorescent protein
RNA	Ribonucleic acid
SNR	Signal-to-noise ratio
SSA	Stochastic simulation algorithm

List of publications

This thesis is a compilation of the following publications:

I Hakkinen, A., Kandhavelu, M., Garasto, S., and Ribeiro, A. S., "Estimation of fluorescence-tagged RNA numbers from spot intensities," *Bioinformatics*, vol. 30, no. 8, pp. 1146 – 1153, 2014

II Hakkinen, A. and Ribeiro, A. S., "Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data," *Bioinformatics*, vol. 31, no. 1, pp. 69 – 75, 2015

III Hakkinen, A. and Ribeiro, A. S., "Characterizing rate limiting steps in transcription from RNA production times in live cells," *Bioinformatics*, in press, doi: 10.1093/bioinformatics/btv744, 2015

IV Hakkinen, A., Tran, H., Yli-Harja, O., and Ribeiro, A. S., "Effects of rate-limiting steps in transcription initiation on genetic filter motifs," *PLoS One*, vol. 8, no. 8, p. e70439, 2013

The author of this thesis contributed to the publications as follows. In **Publication I**, the author designed and implemented the methods, contributed in performing the Monte Carlo simulations and analyzing the measurement and simulation data, and drafted the manuscript. In **Publications II** and **III**, the author designed and implemented the methods, performed the Monte Carlo simulations, analyzed the measurement and simulation data, and drafted the manuscript. In **Publication IV**, the author contributed in conceiving the study, designed the Monte Carlo simulations, contributed in performing the simulations, analyzed the data, and drafted the manuscript. **Publication IV** will also appear in Huy Tran's doctoral thesis.

1 Introduction

1.1 Background and motivation

Gene expression is a process fundamental to life. It is used by all known living organisms, ranging from viruses to multicellular organisms, to transform the genetic information stored in the DNA into the macromolecules which have a functional role in cellular processes. This provides a means for an organism to apply the hereditary information accumulated by generations of ancestors to its survival. Accordingly, rate, timing, and location of the expression of the genes, particularly of the essential ones, is critical for the survival of the cells, and their disruption results in cell death.

The process of gene expression is inherently complex (Alberts et al., 2014). First, the DNA is transformed into a messenger RNA, which is used as a template to synthesize proteins. Each of these processes consists of several steps, which are regulated and catalyzed by numerous molecules. Finally, the proteins must assume the appropriate three-dimensional structure to become functional. Further, the numbers of functional proteins in the cells is affected by the fact that RNAs and proteins have limited lifetime, as they will be eventually degraded by the cellular machinery, in order to allow their continuous renewal. This complexity offers various stages for modulating the dynamics of the gene expression. Such modulation not only determines the rate at which the functional proteins are produced, but can also control the level of stochasticity (i.e. variations) in the resulting protein numbers (Ozbudak et al., 2002).

In bacterial gene expression, such stochasticity is known to arises from various sources. A major part of these fluctuations is attributed to the low copy number effects (Elowitz et al., 2002): the molecular levels tend to be low (Guptasarma, 1995; Taniguchi et al., 2010) and most genes exist as a single copy in the genome, with only a few operator site for each regulatory molecule, implying that any fluctuations in their numbers cause relatively large differences. Also, the fact that the macromolecules are believed to move about primarily by diffusion-like processes (i.e. apparently like a random walk) is another source of randomness (Elowitz et al., 1999). Meanwhile, factors extrinsic to gene expression, such as cellular age (Lindner et al., 2008), uneven partitioning of the molecules at cell

division (Huh and Paulsson, 2011), and heterogeneous environmental conditions (Thattai and van Oudenaarden, 2004) also influence the level of diversity in the numbers of the functional proteins.

The implications of stochasticity on gene regulation and on phenotypic variability between cells has been identified both in prokaryotes and eukaryotes, as well as in stochastic models (Arkin et al., 1998; Blake et al., 2003; Elowitz et al., 2002; Kaern et al., 2005; McAdams and Arkin, 1997; Paulsson, 2004; Raser and O'Shea, 2004; Samoilov et al., 2005; Swain et al., 2002; Weinberg et al., 2005). The variations in protein numbers allow the individuals of populations of genetically identical cells (e.g. of common ancestry) to exhibit phenotypic differences, which can result in widely different behaviors (Arkin et al., 1998; Choi et al., 2008; Elowitz et al., 2002; Suel et al., 2006, 2007). In unicellular systems such as bacteria, the stochasticity alone can be responsible for switching between different phenotypes (Acar et al., 2008; Arkin et al., 1998; Suel et al., 2006), even at the level of single-molecule events (Choi et al., 2008). In addition, the stochasticity causes fluctuations in the behavior of each individual cell over the course of its lifetime (Golding et al., 2005), making cell phenotype a dynamic as opposed to a static property. The consequent diversity of behaviors might allow a better chance of survival for populations of cells as a whole, when subject to unpredictable environmental conditions (Acar et al., 2008). On the other hand, excess stochasticity has been hypothesized to be lethal for essential genes and for genes which encode subunits of protein complexes (Fraser et al., 2004), which agrees with the higher mean levels and relatively smaller noise levels of the essential genes (Taniguchi et al., 2010). These findings imply that there is selective pressure to regulate the amount of stochasticity in gene expression.

Meanwhile, genes do not operate independently but function in networks, in which the elements are connected in an intricate manner (Wolf and Arkin, 2003). As the networks are nonlinear circuits featuring feedback and feedforward, they are capable of exhibiting complex behavior (Becskei and Serrano, 2001; Elowitz and Leibler, 2000; Gardner et al., 2000; Paulsson et al., 2000; Samoilov et al., 2002). The building blocks of these networks are genetic motifs, which are recurring elements found in various organisms adapted to perform a specific function in the network (Shen-Orr et al., 2002; Wolf and Arkin, 2003). Common motifs include genetic switches, which can serve as a memory for maintaining a digital state; oscillators, which can be used for keeping time and synchronizing cellular events; and genetic filters, which are useful for demultiplexing signals and computation via genetic logic (Wolf and Arkin, 2003). While the topology of the motif might determine the primary function of the circuit, the dynamical features of the involved genes define the exact response of these circuits, much like the response of an electronic filter will be determined by the component values.

Consequently, in order to understand how the phenotypic diversity of cell populations can be regulated, it is necessary to understand what kind of dynamical

properties do their components, the individual genes, possess; what are the limits of these properties and how are they regulated, not only in a laboratory environment but in live cells; how are the properties affected by the environmental factors, such as the temperature and stress conditions; and how do they change over the course of the cell lifetime.

Advances in developing fluorescent probes has enabled studying gene expression in living cells (Suzuki et al., 2007). Many different probes have been successfully deployed both in prokaryotes (Golding and Cox, 2004; So et al., 2011; Taniguchi et al., 2010; Yu et al., 2006) and in eukaryotes (Chubb et al., 2006; Fusco et al., 2003; Newman et al., 2006; Raj et al., 2006) for this purpose. Also, the behavior of genetic circuits has been studied using both naturally occurring (Choi et al., 2008) and engineered circuits (Becskei and Serrano, 2001; Elowitz and Leibler, 2000; Gardner et al., 2000). Time-lapse imaging of cells expressing these probes is a particularly effective tool for studying dynamical phenomena (Locke and Elowitz, 2009; Young et al., 2012). Interestingly, some of these techniques allow studying the transcription and translation of particular genes of interest at a single molecule resolution over periods of time (Golding and Cox, 2004; Yu et al., 2006). For example, the MS2-GFP-tagging system of Golding and Cox (2004) allows observing the synthesis of individual messenger RNA molecules, while a technique developed by Yu et al. (2006) allows observing the production of individual fluorescent proteins over time.

As single-molecule measurements of the fluorescent probes are based on fluorescence microscopy images (Coelho et al., 2013), the analysis of the data, particularly at large scales, requires computerized methods. Consequently, a demand for computational and statistical methods for the data analysis has emerged (Locke and Elowitz, 2009; Young et al., 2012). Moreover, since gene expression is inherently stochastic, the results must be extracted and interpreted in a statistical manner. For example, quantification of average behaviors poorly characterizes the behavior of diverse cell populations, an extreme example being populations which exhibit bimodal behavior (Acar et al., 2008). Finally, statistical methods might be the only means of estimating features, which cannot be directly measured in live cells (Kandhavelu et al., 2011; So et al., 2011). As such, the availability and performance of such methods is paramount for the success of single-molecule gene expression studies in live cells.

There are some features, which the methods for data extraction must possess. First, it is important that the methods make no assumptions (i.e. model) of the phenomena being studied, or that the model is sufficiently vague. As such, the methods should be designed with the particular application in mind in order to avoid biasing the results. Designing such methods might be non-trivial, and requires knowledge of the application. For example, some degree of regularization is necessary in order to avoid too complex models, but the regularization must be performed in the appropriate space (e.g. state space versus time) or the results

will be biased. Next, it is preferred that the methods are automatic, as for large collections of data, manually assisted quantification is excessively laborious, and human intervention might introduce superfluous biases or variations which hamper the objective comparison and reproducibility of the results. Further, in some cases, the methods must be robust to outliers. For example, some objects might be missing due to being badly focused or not detected by the earlier stages of the analysis process. Finally, it is preferred that the same methods work for a wide range of conditions and are sensitive to continuous changes, which is necessary for performing differential analysis (i.e. to draw conclusions from the changes in behavior) of the measurement data.

1.2 Objectives

This thesis focuses on methods for extracting information from single-molecule live cell measurements for the purposes of characterizing transcriptional dynamics, and understanding the role of the quantified dynamical properties on the behavior of genetic networks. The focus is on transcriptional dynamics (as opposed to e.g. fluctuations in protein numbers), as, in bacteria, most regulation occurs at the transcriptional level and evidence suggests that the transcriptional kinetics varies widely between promoters and different conditions, suggesting that it is likely to have profound effects on the cell behavior. Here, *Escherichia coli* is used as a model organism, as there is a wealth of information available, the mechanisms tend to be simpler than in eukaryotes, and these bacteria are relatively easy and inexpensive to grow and culture in a laboratory setting.

The methods developed in this thesis aim to generate novel and more accurate quantification of the results from the measurement data, and to allow their statistical analysis in terms of determining confidence in the estimated quantities and performing hypothesis testing on the results. In each stage, the methods are validated using both mathematical methods as well as with novel measurement data of single-RNA dynamics in live *E. coli*. Finally, stochastic modeling of genetic networks is used to study to what extent are the changes in the quantified features of transcriptional dynamics of the component genes reflected on the behavior of genetic networks.

Finally, with minor modifications, the methods are expected to be adaptable to other settings, such as for other fluorescent-tagged molecules, to eukaryotic cells, and for analyzing translational dynamics. As the techniques of fluorescent tagging and microscopy and further evolve, the methods are expected to enjoy wider applicability.

The following objectives were set:

I Quantify GFP-tagged molecular abundances based on fluorescence intensity data. Where applicable, this method should exploit temporal correlations

for greater accuracy and feature robustness against missing objects.

II Estimate the dynamical parameters of subprocesses of transcription, under a wide range of conditions and genes. A statistical method is required, as these subprocesses cannot be directly quantified in live cells.

III Study the effects of changes in transcriptional dynamics, both of the rate and of the shape (stochasticity, skewness, etc.), on small genetic motifs. Here, the focus is on two motifs, one of which performs filtering in the amplitude and the other in the frequency domain.

This thesis accomplishes the objectives as follows: **Objective I** was completed in **Publication I** for static images and in **Publication II** for measurements with temporal information, **Objective II** was completed in **Publication III**, and **Objective III** in **Publication IV**.

1.3 Thesis outline

The thesis is organized as follows. Chapter 2 describes the biological background and the measurement techniques used in observing the RNA production events in live *E. coli* over time. Chapter 3 focuses on modeling transcription and the related forward problems, such as analysis and simulation of the models. Chapter 4 discusses about the general statistical methods on which the developed methods are built on. As opposed to the previous chapter, such methods are used to solve inverse problems, that is, finding an appropriate model for the measurement data. Finally, conclusions and discussion is found in Chapter 5.

2 Biological background and methods

2.1 Gene expression

2.1.1 Central dogma of molecular biology

In all living organisms, the genetic information is encoded in three classes of polymers: DNA, RNA, and proteins. Each of the classes has a different role, which is reflected on their structure and properties. Genetic transfers allow transformations between the polymers of different classes. The basic transfers, DNA replication, transcription (transformation of DNA to RNA), and translation (transformation of RNA to a protein), occur in most cells and are critical for maintaining life. Meanwhile, special transfers, such as reverse transcription, RNA replication, and direct translation of DNA to proteins are known to occur only under special circumstances, such as in viruses or in controlled laboratory conditions. (Alberts et al., 2014; Crick, 1970)

Deoxyribonucleic acid (DNA) is a polymer, which is used for long term storage of the genetic information. It consists of two strands, which run in opposite directions and are complementary to each other. The two strands consist of sequence of nucleotides, which encode the genetic information, and are linked via bonds between the complementary nucleotides. Due to the asymmetric nature of the DNA, it must be constructed in a specific direction ($5'$- to $3'$-end). A typical bacterial genome is featured in a circular DNA segment, which is a few million nucleotides in length. Meanwhile, higher organisms, such as human, have their genome organized in several DNA segments, totaling billions of nucleotides in length. Different regions of DNA encode different types of information. The regions, which encode for functional molecules, such as a protein or a functional RNAs, are called genes. In genes, the regions actually coding the functional polymer are flanked by regions which are responsible of controlling the expression of the gene, such as the promoter region found in the upstream of the coding sequence. The DNA segments are long lived, and are only replicated at cell division. The replication is performed by an enzyme called DNA polymerase, and results in two copies of DNA, each with a new and an old strand.

Meanwhile, ribonucleic acid (RNA) is primarily used for transmission of information from DNA to proteins. Such RNA molecules are called messenger RNAs. Some RNA segments also have functional roles, such as the transfer RNA, which participates in translation, or interfering RNAs, which are involved in gene regulation much like proteins. Structurally, RNA molecules differ from DNA in that the former are constructed using less stable bonds. Due to this, RNAs are short lived, typically with an average lifetime of few minutes (Bernstein et al., 2002; Taniguchi et al., 2010), after which they are degraded by the cellular machinery. The short lifetime allows quick response in regulating the RNA numbers. In addition, the RNA molecules are much shorter, ranging from tens to thousands of nucleotides in length, and are typically present in a single-stranded form.

Transcription is the process of transforming a segment of DNA into an RNA. Transcription is performed by an RNA polymerase, which recognizes the promoter region located at the 5′-end of a gene in the DNA. This process is regulated by transcription factors, which are typically proteins (or e.g. interfering RNA). For example, a transcription factor could bind near the initiation sites, preventing the polymerase from initiating transcription (Schlax et al., 1995). At a specific start site, the DNA is unwound, and RNA is synthesized complementary to one of the DNA strands. The synthesis terminates when the polymerase reaches a stop sequence, causing the newly synthesized RNA to be released. Transcription in *Escherichia coli* is described in more detail in Section 2.1.2.

Meanwhile, proteins are polypeptides, which typically have a single functional role in the cell. Such role can involve regulating cellular processes, such as transcription and translation, catalysis of metabolic reactions, intra- or extracellular signaling, or formation of cellular structures. Structurally, proteins consists of amino acids, each of which is encoded by a sequence of three nucleotides (codon) in the DNA (and RNA). Proteins are synthesized via translation. The synthesis is performed by protein-RNA complexes called ribosomes, which use messenger RNA as a template. Translation starts at a specific site in messenger RNA, where the ribosome subunits are joined, after which the ribosome then elongates along the template synthesizing the protein. The synthesis stops at a specific stop codon. After synthesis, proteins must fold to an appropriate three-dimensional structure to become functional.

The genetic transfer mechanisms in eukaryotic cells feature additional complexities which were omitted from the above description, such as post-processing of RNAs and proteins after their synthesis. Also, in eukaryotes, transcription occurs in the nucleus, while translation occurs in the cytoplasm, which means that the two processes are spatially decoupled. Meanwhile, transcription and translation can occur in parallel in prokaryotes, allowing the translation to begin as soon as the ribosome binding in the beginning of the messenger RNA site has been synthesized.

The number of functional proteins in the cells can be modulated at nearly any stage

of the gene expression process. In addition to their synthesis, both the messenger RNA and proteins are subject to degradation, which influence the abundances of the functional proteins. Typically, most regulation of the protein numbers occurs at transcriptional, post-transcriptional, translational, or post-translational level (e.g. protein folding), transcriptional regulation being the most important, as it occurs early in the gene expression process.

2.1.2 Transcription in *Escherichia coli*

In bacteria, RNA numbers are mostly controlled during transcription. This is suggested by the observations that most regulatory molecules act by modulating the process of transcription initiation (McClure, 1985), and by the apparent lack of correlation between the RNA numbers and their degradation rates (Bernstein et al., 2002). Importantly, the RNA number dynamics also control the level and stochasticity of functional protein numbers. Regulating gene expression at the transcriptional level provides some advantages over translational regulation, such as minimizing the production of superfluous intermediate molecules (Alberts et al., 2014). Meanwhile, transcriptional dynamics is known to vary widely between different promoters, and for the same promoter when under different environmental conditions (Chong et al., 2014; Golding et al., 2005; Kandhavelu et al., 2012a; Muthukrishnan et al., 2012, 2014; Yu et al., 2006), indicating that different regulatory patterns are of importance.

Transcription consists of three major steps. The first step is the transcription initiation, during which the machinery necessary for RNA synthesis is assembled (McClure, 1985). Next step is elongation, which is the nucleotide-by-nucleotide synthesis of the messenger RNA according to the DNA template (Uptain et al., 1997). Finally, the synthesis is terminated, resulting in the release of the newly created messenger RNA and the transcription complex assembly (Nudler and Gottesman, 2002). The process is illustrated in more detail in Figure 2.1.

The process is performed by an enzyme called RNA polymerase. An *E. coli* polymerase holoenzyme consists of a core unit, which is involved throughout the transcription, and a sigma factor which is involved in recognizing specific genes and initiating their transcription. Most *E. coli* genes are recognized by σ^{70}, but there are other sigma factors, which recognize e.g. genes expressed under adverse conditions.

Transcription is initiated at specific sites of the DNA called promoters, which are located at the upstream of the coding region of the gene. The initiation begins with the RNA polymerase holoenzyme recognizing the promoter, resulting in a formation of a transcriptionally inactive complex called “closed complex”. This is followed by the isomerization of the closed complex to form an transcriptionally active complex, dubbed the “open complex”. Finally, the promoter is cleared, and the synthesis of messenger RNA begins. (McClure, 1985).

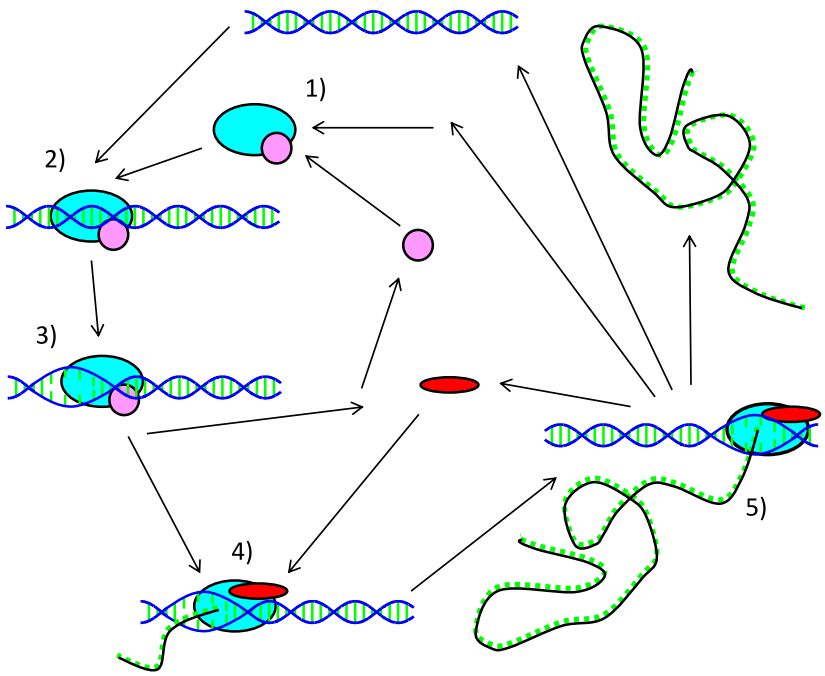

Figure 2.1: Illustration of the transcription process in *E. coli*. 1) The RNA polymerase core enzyme (cyan) and the sigma factor (purple) form the holoenzyme; 2) the holoenzyme binds specifically to the DNA (blue) at the transcription start site; 3) at the start site, the DNA is unwound, and the open complex is formed; 4) the sigma factor is released, while the nusA elongation factor (red) participates in forming the elongation complex, and the elongation begins; 5) at the termination site elongation is terminated, and the polymerase core enzyme, elongation factors, and the newly produced messenger RNA (black) are released.

The primary means by which the RNA polymerase core unit reaches the promoter is three-dimensional diffusion (Wang et al., 2013). The core unit can also weakly bind to the DNA nonspecifically and transfer rapidly between neighboring DNA segments diffusing along the template (Kabata et al., 1993) or by hopping. Next, the core unit forms the polymerase holoenzyme with a sigma factor, which is needed to bind specifically to the promoters of certain genes (Burgess et al., 1969). For example, the $E\sigma^{70}$ holoenzyme recognizes the promoter using specific sequences located at around 10 and 35 nucleotide upstream of the site where the transcription begins. The importance of these sites has been verified by mutations (Walter et al., 1967). Some polymerases, like the bacteriophage T7 RNA polymerase, do not require a sigma factor for specific DNA binding (Bai et al., 2006).

After the RNA polymerase holoenzyme has recognized the promoter, the assembly isomerizes to form a transcriptionally active complex (Saecker et al., 2011). In this process, first both the holoenzyme and the DNA must undergo conformational changes. This is followed by the opening of around 13 basepairs of DNA from -10 (i.e. 10 nucleotides upstream of the transcription start site) to slightly past the transcription start site, by breaking the bonds between the two DNA strands. This results in the creation of the initiation "bubble" and an unstable open complex (Gries et al., 2010). Next, the $+1$ nucleotide of template DNA strand is placed in the active site of the RNA polymerase, while the non-template strand is placed in the binding track, which stabilizes the open complex. On some promoters, the effects of the above steps may be reversed as long as the complex has not yet fully stabilized. However, on strong promoters the steps tend to be essentially irreversible (Record et al., 1996).

Once the fully stable promoter open complex has been formed, the assembly escapes the promoter site (Hsu, 2002) and enters elongation, the productive phase of the RNA synthesis. Subsequently, another RNA polymerase can enter the promoter region and initiate the next transcription event. In the promoter escape process, the sigma factor is released. However, at this stage of transcription, there is a chance that only a short, around 12-nucleotide RNA sequence, is produced, and the transcription process is aborted. This event is known as abortive initiation (Goldman et al., 2009).

Afterwards, the nusA elongation factor assists the RNA polymerase core enzyme in forming the elongation complex. The elongation complex elongates along the DNA strand, synthesizing the RNA product nucleotide-by-nucleotide, until finding a termination site. The RNA synthesis uses one of the DNA strands as a template, creating a sequence of complementary nucleotides, which results in a single-stranded messenger RNA molecule. The building blocks of the RNA are nucleoside triphosphates (NTPs), which contain a nucleotide, which is used for encoding the messenger RNA, and two extra phosphate bonds, which are used to deliver the energy to drive the elongation process. The polymerase moves with single-nucleotide steps, and each step is kinetically a competition between the addition of a nucleotide, a pause, an arrest, or termination of transcription (von Hippel, 1998). Pauses are transient states during which the complex cannot elongate, while arrests are longer stoppages that require the assistance of specific factors in order to resume the elongation process (Bai et al., 2006).

On termination, the RNA polymerase core enzyme, the elongation factor, and the newly synthesized RNA are released. Transcription termination can be triggered by either a specific DNA sequence, in which case the process is called intrinsic termination, or can be triggered by protein factors (Richardson, 2002). In intrinsic termination, the DNA encodes for a sequence, which causes the newly synthesized RNA to form a hairpin loop. Such a hairpin causes the destabilization of the elongation complex. (Nudler and Gottesman, 2002) Meanwhile, in rho-dependent

termination, the rho protein acts as a factor and assists in separating the DNA, messenger RNA, and the elongation complex (Richardson, 2002).

2.2 Single-molecule RNA measurements

There are a few techniques, such as fluorescence in situ hybridization (So et al., 2011; Taniguchi et al., 2010) and RNA sequencing (Croucher and Thomson, 2010), which allow measuring both the transcript levels and their population variance in live bacteria. However, such methods require the cells to be lysed for recording a measurement, implying that they cannot be used to study the behavior of individual cells over time. Meanwhile, fluorescent tags can be used in observing molecular events in live cells over time (Suzuki et al., 2007).

2.2.1 MS2-GFP tagging system

The MS2-GFP RNA-tagging system was implemented in *E. coli* by Golding and Cox (2004). The method is based on a technique originally developed to localize RNA particles in yeast (Bertrand et al., 1998; Fusco et al., 2003). The method is unique in that it allows observing the production of the target transcripts in live bacteria at the individual transcript resolution over periods of time. Consequently, it allows studying the process of transcription in the absence of other processes which influence the RNA numbers, such as RNA degradation (Taniguchi et al., 2010) and their dilution by the cell division (Huh and Paulsson, 2011).

The system consists of two elements: a green fluorescent protein (GFP) reporter fused to the bacteriophage MS2 coat protein, which allows the fusion protein to bind specifically, and a target RNA containing tandem repeats of the MS2 binding sites. The elements are illustrated in Figure 2.2.

More specifically, the reporter gene encodes for a fusion protein of MS2d, a tandem dimer mutant of the wild type MS2 (Peabody and Lim, 1996), and GFPmut3 (Cormack et al., 1996), an optimized variant of GFP. The reporter gene is hosted in a medium-copy plasmid and is expressed constitutively, which implies that the MS2-GFPs are abundant in the cells at all times. Meanwhile, the MS2 coat protein features high specificity and high binding affinity to the MS2 hairpin sequences (Johansson et al., 1998). Consequently, any transcribed MS2 binding sites will be shortly occupied by an MS2-GFP protein, and nearly 100% of the MS2-GFP binding sites are expected to be occupied (Golding et al., 2005) at all times. As the proteins are abundant, no significant change in the background intensity of the cell is expected on RNA production (Golding et al., 2005). These characteristics make the system independent of slow and highly stochastic cellular processes, such as protein production, folding, and chromophore maturation, which limit the applicability of some other reporter systems.

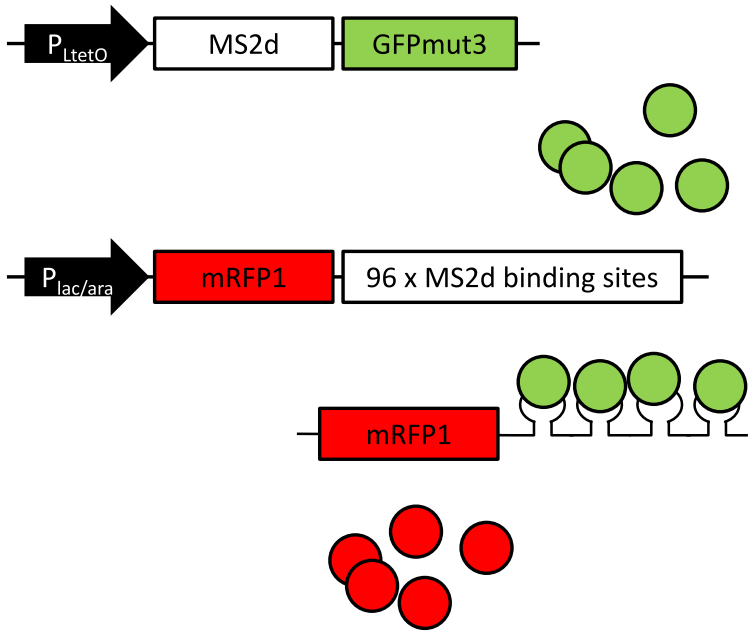

Figure 2.2: Illustration of the components of the MS2-GFP RNA-tagging system (Golding et al., 2005). The reporter, here controlled by the LtetO promoter, encodes for the MS2-GFP fusion protein. Meanwhile, the target RNA, controlled by lac/ara promoter, encodes for mRFP1 and 96 binding sites for the MS2 coat protein. After a target transcript is produced, the abundant MS2-GFPs bind to it, allowing it to be visualized as a cluster of GFPs. Finally, the target RNA is translated, which can be detected by observing the presence of RFP in the cells.

Meanwhile, the target gene is hosted in a single-copy F-based vector, which allows observing transcription events at the resolution of a single gene. The gene encodes for MS2 hairpin repeats in either at the $5'$- or $3'$-end of the transcript, which allow the transcript to be tagged by the MS2-GFP fusion proteins. The original construct contains 96 of such sites (Golding and Cox, 2004). After the binding sites have been transcribed, they are rapidly occupied by the MS2-GFPs, even prior to the completion of transcription (Golding et al., 2005). This results in an intense, well resolved spot in the cells, which can be visualized using fluorescence microscopy, as exemplified in Figure 2.3. Meanwhile, the binding of the MS2-GFPs to the target RNA appears to prevent RNA-degrading enzymes from degrading it, essentially immortalizing the target RNA (Golding and Cox, 2004; Muthukrishnan et al., 2012). This implies that the RNA quantification is not affected by other processes such as RNA degradation.

In addition, the target gene also encodes for a monomeric red fluorescent protein (RFP), mRFP1 (Campbell et al., 2002). While it might not be possible to identify

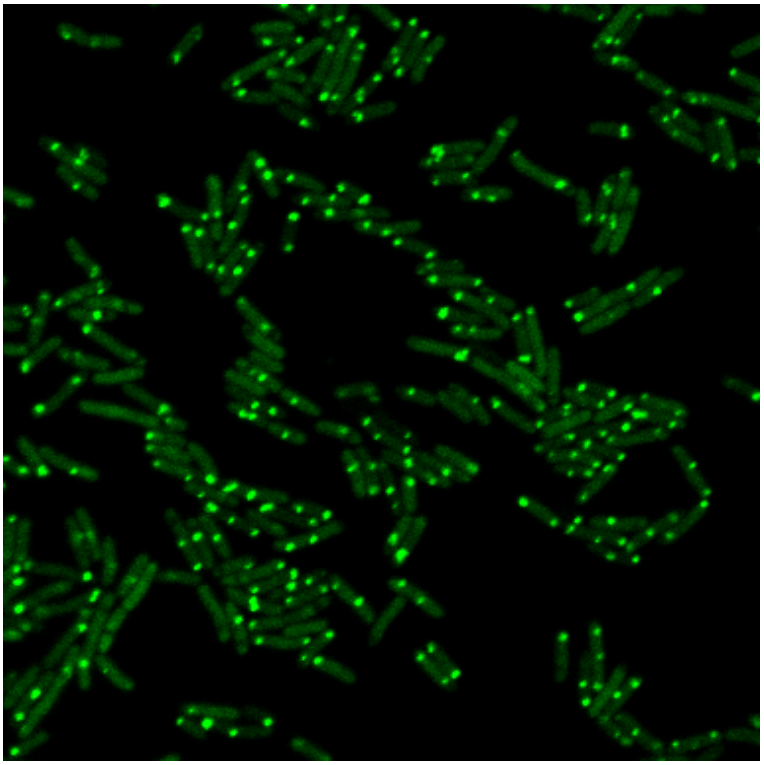

Figure 2.3: Live *E. coli* expressing MS2-GFP and the target RNAs visualized using fluorescence confocal microscopy. The MS2-GFPs are abundant and uniformly distributed in the cells, making the cell backgrounds visible. The produced target RNAs contain 48 binding sites for the MS2-GFPs, making them to appear as bright green spots.

individual proteins in the images, the RFP signal allows correlating the protein abundance with that of the target RNA (Golding et al., 2005). This allows validating the changes in expression levels using independent methods, such as by using a microplate reader. Following fluorescence microscopy imaging, the number of RNA molecules in each cell or in each fluorescent RNA spot can be estimated based on their intensity (Golding et al., 2005). Such process involves measuring the intensity emitted by a single tagged RNA, and using it as a normalizing factor to scale the intensities of the fluorescent spots.

Golding originally constructed a few constructs with different target promoters, such as the lac, lac/ara-1, and bacteriophage λ RM promoter (Golding and Cox, 2004; Golding et al., 2005). More recently, additional variants were engineered, which contain other target promoters (Makela et al., 2013; Muthukrishnan et al., 2012). Such libraries of constructs allow understanding gene expression dynamics

in a wider scale. Also, a variant with lac/ara-1 promoter and a target RNA containing only 48 binding sites has been engineered (Hakkinen et al., 2014), which allows determining if the RNA quantification limits the usability of the system.

2.2.2 Image processing

Performing large-scale analysis of fluorescence microscopy images necessitates methods which can extract the relevant information with minimal human intervention. As opposed to manual analysis of the images, the automatic methods are also advantageous in that they produce objective results, as they lack variations resulting from the choices of different experts.

For this work, the following steps for the analysis are used. First, the cells are segmented using one of the two algorithms: a semi-automatic method described in Kandhavelu et al. (2012b) or an automatic method proposed in Chowdhury et al. (2013). Next, the fluorescence in the cell background must be estimated and subtracted, as the GFPs are highly abundant in the cells, resulting in strong uniform fluorescent cell backgrounds. Afterwards, the fluorescent RNA spots are segmented using either a method from Ruusuvuori et al. (2010) or the one proposed in Hakkinen et al. (2014). Finally, the volume of the spots above the estimated cell background is integrated to obtain the background-corrected total fluorescence for each cell at each time moment. An example of the results at different stages of this process is shown in Figure 2.4.

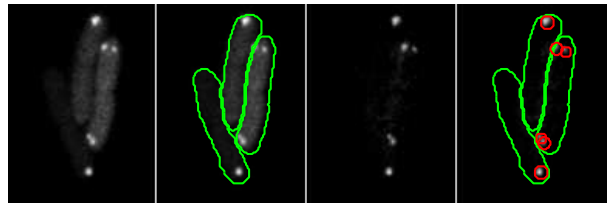

Figure 2.4: RNA spot intensity extraction from confocal microscope images. From left to right: original intensity image with three cells and six RNA spots; borders of the segmented cells superimposed on the previous image; intensity after subtracting the estimated cell backgrounds; cell borders and spot isolines superimposed on the previous image.

The first set of methods has been successfully used in various publications (Kandhavelu et al., 2012b,a; Makela et al., 2013; Muthukrishnan et al., 2012), but result in higher noise in the spot intensities (Hakkinen et al., 2014), and consequently, lower confidence in the extracted RNA numbers or RNA production times. The

latter set of methods were used to analyze individual frames of cell populations in **Publication I**, where the temporal information cannot be exploited, and the accurate quantification of the spot intensities is critical. Meanwhile, the temporal measurements in **Publications I**, **II**, and **III** use the former set of methods.

Cell segmentation generates a labeled image in which each pixel in the fluorescence image is associated either with image background, or with one of the cells. The cell segmentation method proposed in Kandhavelu et al. (2012b) requires the user to manually divide the image in separate regions occupied by each of the selected cells, preventing clusters of cells to be mistaken as single cells. After this, the locations, dimensions, and orientations of the cells are found using principal component analysis (PCA) (Hotelling, 1933; Pearson, 1901). For this, it is assumed that the fluorescence inside each cell is uniformly distributed in some rectangle, which is used to scale the intensity covariance appropriately. Essentially, this produces the center of mass and the major and minor axes of the box that best represents the cell.

The segmentation of highly clustered cells requires more sophisticated methods, since large clusters make the segmentation method of Kandhavelu et al. (2012b) laborious. Therefore, more recently, an automatic method proposed in Chowdhury et al. (2013) was adapted for this process. The method employs an image denoising filter (Dabov et al., 2007), blockwise thresholding with Otsu's method (Otsu, 1979) to separate the cell clusters from the background, and multiscale morphological edge detection (Chowdhury et al., 2013) to segment the clusters into an initial set of candidate segments. Afterwards, the segmentation is refined by iteratively learning features, such as dimensions of the objects, and using this information to split and merge the segments while increasing the likelihood of the segmentation given the estimated feature distributions (Chowdhury et al., 2013).

In order to quantify the production of RNA over time, the cells must be tracked in the sets of images. The cell lineages are established by computing the overlapping areas of the cell segments in consecutive frames after globally aligning the images (Hakkinen et al., 2013b). In each frame, a cell segment is associated with that with the most overlap in the previous frame. If there are multiple such cells, it is taken to mean that the cell has divided. Due to the frequent imaging, there is little movement between the consecutive frames and the method tends to work well, provided that the segmentation is accurate.

In Hakkinen et al. (2014), the cell background is estimated by fitting the intensity of each cell with a surface, which is a quadratic polynomial of the distance from the cell border (obtained from the segmentation), in least-deviations sense. The least-deviation criterion ensures that the surface fit disregards the RNA spots, assuming that they occupy a minority of the cell area. Such assumption is reasonable as long as the number of RNA spots is low (e.g. less than 10). The background estimated with this method is illustrated in the middle panel of Figure 2.5.

The background model in Kandhavelu et al. (2012b) is a constant surface, and

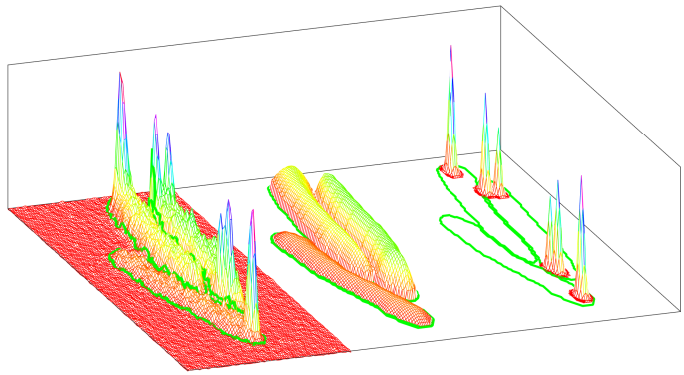

Figure 2.5: Estimated cell background and foreground surfaces. From left to right: surface plot of the original intensity images with three cells and six RNA spots (same data as in Figure 2.4); the estimated cell backgrounds; and the estimated fluorescence spots. The green and red curves represent the cell borders and the spot isolines, respectively.

consequently, it is possible to subtract it after the spot detection, provided that the spot detection method is shift-invariant in the intensity level. The height of this surface is the average intensity of the cell outside the RNA spots, that is, the process corresponds to fitting a constant surface in least-squares sense to the cell background disregarding the spots. As the spots, which are outliers in the intensity space, are disregarded, the robustness of the least-deviations criterion is not required. In such case, a least-squares estimator is expected to be a more accurate, provided that the errors are approximately normal.

The spot detection method from Ruusuvuori et al. (2010) operates as follows. First, kernel density estimation (KDE) (Parzen, 1962; Rosenblatt, 1956) is used to estimate the probability density of intensity values in some local neighborhood $N(i, j)$ around the pixel at (i, j). Next, the likelihood that the intensity $I(i, j)$ of the pixel at (i, j) comes from the distribution specified by its neighbors is computed. Finally, the likelihoods are thresholded to obtain a binary image of the spots. Here, a circular window for the neighborhood and a Gaussian kernel were used, and the threshold was found using Otsu's method (Otsu, 1979). The neighborhood radius and the KDE bandwidth were tuned manually.

Meanwhile, the spot detection method from Hakkinen et al. (2014) operates as follows: After subtracting the cell background, the remaining intensity is fit with a set of multidimensional Gaussian functions, in least-deviations sense, with decreasing heights until the heights are in the 99% confidence interval of the background noise. The background noise is determined by assuming a normal distribution and estimating its variance by computing the median absolute deviation. The Gaussian functions are taken to represent spots, the volume under each representing the total spot intensity. Meanwhile, the volume under the whole

foreground surface is taken to represent the total cell intensity. The results of this method are exemplified in the right panel of Figure 2.5.

3 Modeling

3.1 Stochastic modeling of chemical reactions

The most accurate way of modeling chemical systems is to perform molecular dynamics simulations. In such simulations, the positions and velocities of each molecule in the system must be tracked over time. In contrast to this, stochastic chemical kinetics makes some further assumptions which allow modeling the positions and velocities using random variables instead.

The stochastic approach correctly accounts for the discrete nature of molecule numbers and the correlations between them (McQuarrie, 1967). The formulation presented here builds on the works of Gillespie (1976, 1977a,b, 1992, 2007). There are also approximate stochastic methods, which feature stochasticity but with an incorrect shape, such as the chemical Langevin equation (Gillespie, 2000), which are not covered here.

3.1.1 Stochastic chemical kinetics

Consider a system of n chemical species, here denoted by $S_1, \cdots, S_n$, which can interact through m chemical reaction channels, denoted by $R_1, \cdots, R_m$. The system state is represented by the state vector $\mathbf{X}(t) \doteq (\, X_1(t), \cdots, X_n(t)\,)^*$ of random variables, where $X_i(t)$ denotes the number of molecules of chemical species S_i in the system at time t. Further, the system is assumed to be at some known state $\mathbf{X}(t_0) = \mathbf{x}_0$ at some point in time t_0; equivalently, one could assume some non-degenerate probability distribution defining the state at t_0, but the former is used here for convenience. As the system is probabilistic, the fundamental question is to determine the how the probability distribution of $\mathbf{X}(t)$ evolves over time, given the initial condition.

The system is assumed to be of constant volume, be in thermal equilibrium at a constant temperature, and be well-stirred. The last assumption is guaranteed, for example, by the fact that the majority of molecular collisions are nonreactive, which is true for dilute gas systems but also tends to hold more generally. Instead of modeling the positions and velocities of each individual molecule in the system, the above assumptions allow their positions and velocities to be modeled as random variables. Namely, the above implies that the positions of the molecules

are uniformly distributed in the volume, their velocities are Maxwell-Boltzmann distributed, and that these distributions do not change over time.

The changes in the state vector $\mathbf{X}(t)$ are induced by the m chemical reaction channels. Each reaction channel R_j has an associated state change vector $\boldsymbol{\nu}_j \doteq (\,\nu_{1,j}, \cdots, \nu_{n,j}\,)^*$, where $\nu_{i,j} \doteq \nu^+_{i,j} - \nu^-_{i,j}$ is the change in the number of molecules of the species S_i induced by the reaction channel R_j. Here, for convenience, $\nu^-_{i,j}$ and $\nu^+_{i,j}$ represent the number of molecules consumed and produced by the reaction, respectively. Also, each reaction channel R_j features an associated propensity function $a_j(\mathbf{x})$, which determines how frequently the reaction occurs. Specifically, $a_j(\mathbf{x})\ \partial t$ represents the probability that exactly one reaction of the channel R_j occurs in the infinitesimal time window $[t, t + \partial t)$, given that the system is in the state $\mathbf{X}(t) = \mathbf{x}$ at time t. The fact that this probability can be expressed in such a form is the fundamental premise of stochastic chemical kinetics.

The validity of the fundamental premise depends on the physics of the particular reaction. If the reaction channel R_j is an unimolecular reaction, that is, of the form $S_j \to \cdots$, its physics are expected to be internal to a single molecule and dictated by some quantum mechanical phenomenon, analogous to nuclear decay (Gillespie, 1992). Therefore, it is expected that the probability of a particular molecule of species S_j to react is some constant c_j per unit time. As there are x_j molecules of such species in the system, it follows that the probability that exactly one of these reactions occurs in the infinitesimal time window $[t, t + \partial t)$ is $a_j(\mathbf{x})\ \partial t = c_j\, x_j\ \partial t$.

Meanwhile, if R_j is bimolecular, that is, of the form $S_i + S_j \to \cdots$, the occurrence of a reaction is a result of a collision between two molecules. Arguments from the kinetic molecular theory combined with the assumption that the system is well-stirred imply that, again, the appropriate form is available (see e.g. Gillespie, 1992 for the details). Namely, the probability that a specific pair of S_i and S_j molecules collide and react according to R_j in the time window $[t, t + \partial t)$ is given by $c_j\ \partial t$, where c_j is some constant with respect to time. This suggests that the propensity function can be expressed as:

$$a_j(\mathbf{x}) = \begin{cases} c_j\, x_i\, x_j & \text{, for } i \neq j \\ c_j\, \frac{x_i\,(x_i - 1)}{2} & \text{, for } i = j \end{cases}$$

The implied physics of the higher order reactions, that is, reactions of the form $S_i + S_j + S_k + \cdots \to \cdots$, is unclear. One could argue that the such reactions never appear as elementary reactions, but are composed of a set of lower order reactions. In such case, the justification is not necessary, and the modeling should be performed using the lower order reactions, unless an approximation is acceptable. Regardless, one expects that a reaction of the form:

$$\nu^-_1\, S_1 + \cdots + \nu^-_n\, S_n \xrightarrow{c_j} \cdots$$

has a propensity function of the form:

$$a_j(\mathbf{x}) = c_j \prod_{i=1}^{n} \binom{x_i}{\nu_{i,j}^-} = c_j \prod_{i=1}^{n} \frac{x_i\,(x_i-1)\cdots(x_i-\nu_{i,j}^-+1)}{\nu_{i,j}^-\,(\nu_{i,j}^--1)\cdots 1}$$

where c_j is some constant, and the product of binomial coefficients represents the number of combinations of the reacting molecules. Such form is in accordance with the two above forms, and satisfies the fundamental premise.

Now, let $P(\mathbf{x},t) \doteq \mathbb{P}[\,\mathbf{X}(t)=\mathbf{x}\,|\,\mathbf{X}(t_0)=\mathbf{x}_0\,]$ denote the density of system state, given the initial condition. Given the premise, it is possible to write the density after an infinitesimal time step $P(\mathbf{x},t+\partial t)$ in terms of the current density $P(\mathbf{x},t)$, from which it can be concluded that the density must evolve as follows:

$$\frac{\partial}{\partial t}\,P(\mathbf{x},t) = \sum_{j=1}^{m} a_j(\mathbf{x}-\boldsymbol{\nu}_j)\,P(\mathbf{x}-\boldsymbol{\nu}_j,t) - \sum_{j=1}^{m} a_j(\mathbf{x})\,P(\mathbf{x},t)$$

which is called the chemical master equation (CME). As the CME is just a system of linear ordinary differential equations (ODEs), it completely determines the function $P(\mathbf{x},t)$, provided that some initial condition is given. A complication, which makes obtaining direct solutions from CME hard, is that the system is potentially of infinite dimension, as there is one variable for each combination of numbers of the molecules.

Taking the expectation of both sides of the master equation yields:

$$\frac{\partial}{\partial t}\,\mathbb{E}[\,\mathbf{X}(t)\,] = \sum_{j=1}^{m} \boldsymbol{\nu}_j\;\mathbb{E}[\,a_j(\mathbf{X}(t))\,]$$

which again is a system of ODEs, now of finite dimension n. If all the reactions are of zeroth or first order, the ODEs are linear, in which case the expectations can be solved in closed form. However, if there is a single bimolecular reaction, an expectation on the right-hand side features moments of $\mathbf{X}$ higher than that of the left-hand side. Consequently, determining the moments of a system with bimolecular (or higher order) reactions would also require solving a system of ODEs of infinite dimension.

Meanwhile, if $\mathbf{X}(t)$ is assumed to be deterministic, then $\mathbb{E}[\,\mathbf{X}(t)\,] = \mathbf{X}(t)$, and:

$$\frac{\partial}{\partial t}\,\mathbf{X}(t) = \sum_{j=1}^{m} \boldsymbol{\nu}_j\,a_j(\mathbf{X}(t))$$

which also governs the state-evolution of the corresponding system of reaction rate equations, which are used in deterministic chemical kinetics (see e.g. Iglesias and Ingalls, 2009). Also, it can be noted from above that when all the reactions are of zeroth or first order, the expectation of the stochastic system equals that of

the deterministic. However, this does not hold when reactions of higher order are presents, demonstrating that stochastic kinetics of bimolecular reactions features intricate dynamics even in the mean levels, which are only captured when the low copy numbers are properly accounted for.

Another link between the stochastic and deterministic kinetics is that the deterministic formulation can be obtained as the infinite molecule limit solution of the CME, provided that the concentrations of the molecules remain constant (Gillespie, 2009; Kurtz, 1972). This demonstrates the fact that the discrepancies between the two strategies arise from the effects of finite (low) copy numbers, vanishing in the limit of heterogeneous system of infinite size.

3.2 Monte Carlo methods

Monte Carlo (MC) methods are algorithms, which employ random sampling to obtain numerical estimates (Van Trees et al., 2013). Such methods are used, for example, for stochastic optimization, numerical integration, and generating random numbers from complicated probability distributions; or any application, where the state space is too large for thorough exploration, often encountered in combinatorial or high dimensional problems.

For many problems, the central limit theorem (CLT) (Lehmann and Casella, 1998) and the continuous mapping theorem (Lehmann and Casella, 1998; Mann and Wald, 1943) together guarantee that the result from a set of independent MC simulations can be made asymptotically correct on average, and the standard deviation grows like $1/\sqrt{n}$, where n is the number of simulations. This implies that MC simulations can be used to obtain results with arbitrary accuracy, which is determined by the number of independent simulations.

3.2.1 Generating random numbers

Performing large-scale MC simulations requires vast amounts of quality random numbers. If the statistical properties of these random numbers have defects, they can be propagated to the results.

True random numbers are based on some physical phenomenon, which is expected to be random, such as nuclear decay. However, the rate at which such numbers can be harvested is limited and requires special hardware. Meanwhile, pseudorandom numbers are numbers generated using a deterministic algorithm that appear to be unpredictable. Such numbers are attractive, since they are inexpensive to generate and the sequence of numbers can be reproduced by seeding the generator with equal settings. Fortunately, some pseudorandom number generators (PRNGs) generate sequences of random numbers, which appear random enough, as determined by some statistical tests, and as such, are likely good candidates for the purposes of MC methods. A disadvantage of PRNGs is that they are necessarily periodic,

as they are deterministic and have a finite state. For large MC simulations, the period must be long, or there will be artificial correlations between the generated values.

Standard PRNGs used for systems programming, such as linear congruential generators, are poorly suited for this task, since they have very short periods, and the samples tend to be heavily correlated (Marsaglia, 1968). Moreover, such generators often feature poor choices of design parameters, which corroborate these problems (Park and Miller, 1988).

One PRNG, which is generally considered to be adequate for numerical simulations is Mersenne twister (Matsumoto and Nishimura, 1998). It generates (discrete) uniform 32-bit random numbers with a period of $2^{19937} - 1$. The random numbers are 623-dimensionally equidistributed up to the 32-bit accuracy, and sequences of the numbers are known to pass numerous tests of statistical randomness.

3.2.2 Transforming random numbers

Generally, the random number generators available for digital computers generate uniform random integers. Such integers are easily converted to floating point numbers, which, for most applications, can be regarded as continuous random numbers in the unit interval (i.e. $[0, 1)$). For practical applications, such numbers must be further transformed to the appropriate distribution.

Random variates, whose distributions have a simple (inverse) cumulative distribution functions (CDFs) can be obtained using a technique called inverse transform sampling (Van Trees et al., 2013). Let U be a unit interval continuous uniform random variable, and let $X = F^{-1}(U)$. Now:

$$\mathbb{P}[\, F^{-1}(U) \leq x \,] = \mathbb{P}[\, U \leq F(X) \,] = F(X)$$

so if $F(x)$ is chosen to be the CDF of X, then X must have the appropriate distribution. This is correct, since the CDF is monotonic and right continuous, and as such, can be inverted, at least in a weak sense: $F^{-1}(u) = \inf \{\, x : F(x) \geq u \,\}$, where $\inf \{\cdot\}$ denotes the infimum, that is, the greatest value not greater than any in the set.

For example, as the CDF of the exponential distribution is $F(x) = 1 - \exp(-x \,/\, \mu)$, exponential variates with mean of μ can be generated using:

$$x = -\mu \, \log(\, 1 - u \,)$$

where u is a continuous unit-interval uniform random variate.

Another common transformation method is rejection sampling. In this method, realizations of the random variable X with the probability density function (PDF) $f(x)$ are generated with the aid some of some other random variables Y which have a simpler PDF, for example, an uniform distribution. Let U be a unit interval

continuous uniform random variable, and let Y be an independent random variable with the PDF $g(y)$, such that for some constant M, $f(x) \leq Mg(x)$ for all x. Now, the random variable $X = Y \mid (U < f(Y)/(M\,g(Y)))$ must have the PDF $f(x)$, since:

$$\mathbb{P}[\,U < f(Y)/(M\,g(Y))\,] = \mathbb{E}[\,f(Y)/(M\,g(Y))\,] = 1\,/\,M$$

and:

$$\mathbb{P}[X \leq x] = \frac{\mathbb{P}[\,U < f(Y)/(M\,g(Y)) \wedge Y < x\,]}{\mathbb{P}[\,U < f(Y)/(M\,g(Y))\,]} = F(x)$$

where $F(x)$ is the antiderivative of $f(x)$. Realizations of X can be generated by generating u and y as realizations of U and Y, respectively, and returning $x \leftarrow y$ only when $u < f(y)/(M\,g(y))$. On average, M iterations are required. There are more sophisticated variants of the rejection sampling method, such as the Metropolis-Hastings algorithm (Hastings, 1970).

3.2.3 Stochastic simulation algorithm

As pointed out, obtaining direct solutions from the CME either in closed form or numerically is difficult for arbitrary nonlinear systems. One numerical method which is applicable on arbitrary systems is the stochastic simulation algorithm (SSA). The SSA is a Monte Carlo method for sampling trajectories $\mathbf{x}(t)$ from the CME (Gillespie, 1976, 2007), that is, $\mathbf{x}(t)$ is a realization of $\mathbf{X}(t)$ with the appropriate distribution. As such, it is not an approximate method despite its numerical nature, as it does not make any further assumptions beyond that of the CME, and consequently, is exact in the sense that the generated trajectories have exactly the probability distribution specified by the CME.

The key in generating such trajectories is not the master equation itself, but a related density. This density, denoted here by $p(\tau, \mu)$, is the probability density that the next reaction occurs in the time window $[t+\tau, t+\tau+\partial t)$, and will be the reaction R_μ, given that the system state $\mathbf{X}(t) = \mathbf{x}$ at time t is known. From the fundamental premise of stochastic chemical kinetics, it follows that the density has the form of:

$$p(\tau, \mu) = a_\mu(\mathbf{x})\,\exp(-a_0(\mathbf{x})\,\tau) = \underbrace{\frac{a_\mu(\mathbf{x})}{a_0(\mathbf{x})}}_{p(\mu)}\,\underbrace{a_0(\mathbf{x})\,\exp(-a_0(\mathbf{x})\,\tau)}_{p(\tau)}$$

with:

$$a_0(\mathbf{x}) \doteq \sum_{i=1}^{m} a_j(\mathbf{x})$$

This fact is the mathematical basis for the SSA. As the density can be factored into the two independent parts, it is apparent that τ must be an exponentially distributed random variate with a rate of $a_0(\mathbf{x})$, while μ must be a categorical random variate with probability $a_\mu(\mathbf{x})\,/\,a_0(\mathbf{x})$ for the reaction channel R_μ.

Such random variates can be generated using the following inverse transform sampling scheme:

$$\tau = \frac{-\log(1-u_1)}{a_0(\mathbf{x})}$$
$$\text{find } \mu \text{ such that } \sum_{j=1}^{\mu-1} \leq u_2\, a_0(\mathbf{x}) < \sum_{j=1}^{\mu} a_j(\mathbf{x})$$

where u_1, u_2 independent and identically distributed (iid) uniform random variates in $[0, 1)$.

Following this, given some initial state $\mathbf{X}(t_0) = \mathbf{x}_0$, the evolution of $\mathbf{x}(t)$ can be generated by maintaining the current system state $\mathbf{x}$ and the current time t, and updating them according to the generated numbers. First, let $\mathbf{x} \leftarrow \mathbf{x}_0$ and $t \leftarrow t_0$, which sets the system to the initial state; if the initial state is specified by a distribution, generate realizations of these distributions, accordingly. For the update, first the terms $a_j(\mathbf{x})$ must be updated, as they depend on the current state, and $a_0(\mathbf{x})$ is computed according to its definition. Next, realizations τ and μ are generated as specified above. Finally, the current system state and the current time are updated to reflect the occurrence of the reaction: let $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$ and $t \leftarrow t + \tau$. This process is repeated until a predetermined point condition is reached (e.g. simulation time), or $a_0(\mathbf{x})$ is zero, in which case no reactions can occur and the state $\mathbf{x}$ is frozen.

Here, τ does not represent an approximation step size, which is characteristic to e.g. numerical integration of ODEs. Instead, it indicates when the next reaction occurs. This implies that $\mathbf{x}$ is well defined and remains unchanged in the time window $[t, t + \tau)$, until it jumps to the next state $\mathbf{x} + \boldsymbol{\nu}_\mu$ at time $t + \tau$.

The disadvantage of the SSA is that it can be slow if reactions are occurring frequently. Particularly, if there are reactions occurring at widely different time scales, it might be difficult to gain insight of the behavior of the system as a whole. There are some approximations and generalizations, which can be used to accelerate the simulations by e.g. coalescing the updates for multiple reactions. For example, the τ-leap method (Gillespie, 2001) approximates the SSA by assuming that the propensity functions do not change frequently. Meanwhile, the delayed SSA (Bratsun et al., 2005; Roussel and Zhu, 2006) can be used to coalesce the simulation of chains of reactions, which exhibit non-Markovian dynamics.

3.3 Modeling transcription

Genetic products often exist in low copy numbers (Bernstein et al., 2002; Ghaemmaghami et al., 2003; Guptasarma, 1995; Taniguchi et al., 2010). Consequently, the fluctuations and correlations in their numbers cannot be neglected, as they result in behaviors which are not captured by just their expression levels (Arkin

et al., 1998; Blake et al., 2003; Elowitz et al., 2002; McAdams and Arkin, 1997; Raser and O'Shea, 2004; Samoilov et al., 2005; Weinberg et al., 2005). Moreover, these processes are inherently complex (McClure, 1985), involving steps such as binding and unbinding of various regulatory molecules, assembly of complexes, diffusion of the assembling molecules through a nucleotide chains of various lengths, and maturation and folding of the produced polymers to their appropriate three-dimensional structure (Alberts et al., 2014). The details of these processes control not only the rate at which RNA and proteins can be produced, but also the level of fluctuations and other dynamical features, such as how fast the genes can respond to transient signals.

Models based on stochastic chemical kinetics have been found to successfully capture the features present in the cellular processes (Arkin et al., 1998; Blake et al., 2003; McAdams and Arkin, 1997; Samoilov et al., 2005; Weinberg et al., 2005). The simplest stochastic model for transcription is the model of spontaneous creation of the product:

$$\emptyset \xrightarrow{k_1} M \quad ,$$

where M represents the product, here a messenger RNA, and k_1 is the rate at which it is being produced. This has been shown to well describe the measurements of single-molecule dynamics in live cells under certain conditions, such as the production of RNA in bacterial genes with slow rates (Yu et al., 2006).

The CME corresponding to this model is:

$$\frac{\partial}{\partial t} P(m,t) = k_1 \, P(m-1,t) - k_1 \, P(m,t)$$

whose Z-transform ($X \mapsto \mathbb{E}[z^X]$) is:

$$\frac{\partial}{\partial t} G(z,t) = k_1 (z-1) \, G(z,t)$$

which is an ODE with the solution $G(z,t) = G(z,0) \exp(\, k_1 \, t \, (z-1) \,)$. Here, $G(z,t)$ is the Z-transform of $P(m,t)$, known as the probability generating function from which $P(m,t)$ can be recovered by inverting the Z-transform. Consequently, the total number of produced RNAs at time t is $M(t) - M(0) \sim \mathcal{P}(k_1 \, t)$, which denotes a Poisson distribution with a mean of $k_1 \, t$. For this distribution, both the expected number of produced RNAs $\mathbb{E}[M(t) - M(0)]$ and its variance $\mathbb{C}\mathrm{ov}[M(t) - M(0)]$ are equal to $k_1 \, t$, indicating that both the mean and the fluctuations of the produced RNA numbers vary over time, and are controlled by the kinetic parameter k_1.

Meanwhile, more detailed models of transcription and translation have been proposed (Klumpp and Hwa, 2008; Makela et al., 2011; Ribeiro et al., 2006, 2009; Roussel and Zhu, 2006), which allow studying more fine grained details of the steps involved in these processes. This has been motivated by the fact that not

in all cases are the mean and variance of RNA numbers controlled by a single parameter (Chong et al., 2014; Golding et al., 2005; So et al., 2011; Taniguchi et al., 2010), as suggested by the above model. Such models may account for details, such as the individual steps of formation of the complexes in transcription and translation initiation (McClure, 1985), premature termination (Grundy and Henkin, 2006) of the transcription and translation processes, stepwise elongation nucleotide-by-nucleotide, or codon-by-codon, with features such as arrests (Greive and von Hippel, 2005), pauses (Herbert et al., 2006; Landick, 2006), and editing and backtracking of the polymerase (Greive and von Hippel, 2005).

3.3.1 Sequential model of transcription initiation

Genes, which lack a stringent regulation by the activator and repressor molecules, have their transcriptional dynamics mainly controlled by the interactions between the promoter region and the RNA polymerase (McClure, 1980). This follows from the fact that the chain elongation during transcription tends to be fast, around 80 to 90 nucleotides per second in *Escherichia coli*, (Vogel and Jensen, 1994), resulting in delays ranging from tens of seconds to couple of minutes. In addition, the process of elongation does not limit the throughput of the gene, but has only implications on its response time (latency), provided that there is no RNA polymerase traffic (Klumpp and Hwa, 2008; Rajala et al., 2010).

In vitro measurements of the transcription process suggest that it consists of multiple sequential steps, which occur during the transcription initiation process, and whose rates limit the overall expression of the gene (Lutz and Bujard, 1997; Lutz et al., 2001; McClure, 1980, 1985). The dynamics of these processes are determined by the gene sequence (as encoded by the DNA), and the conditions, such as abundance of RNA polymerases and metabolites (Lutz and Bujard, 1997; Lutz et al., 2001; McClure, 1985). Motivated by these observations, the following kinetic model has been proposed (McClure, 1985):

$$P + R \rightleftharpoons R{\cdot}P_c \rightleftharpoons R{\cdot}P_i \rightleftharpoons R{\cdot}P_o \rightarrow P + R + E$$

where P represents the promoter, R represents an RNA polymerase, and E is the elongation complex. Here, $R{\cdot}P_c$, $R{\cdot}P_i$, and $R{\cdot}P_o$ are intermediate complexes of transcription, called the closed complex, isomerization complex, and the open complex, respectively. Several different steps have been identified in the isomerization process, but only some of them tend to be rate limiting (Saecker et al., 2011). Besides the promoter sequence, the kinetics of each step are expected to be affected by the presence/absence of regulatory factors and DNA supercoiling configurations (Bai et al., 2006).

In this model, the transcription process starts with an RNA polymerase holoenzyme binding to the promoter, provided that it is not yet occupied. When bound, the RNA polymerase forms an unstable promoter closed complex $R{\cdot}P_c$, which is stabilized through an isomerization step $R{\cdot}P_i$, eventually resulting in the open

complex $R{\cdot}P_o$. The backward reactions represent the reversible nature of the steps, which results from the fact that the intermediate complexes have not yet been fully stabilized. Finally, the open complex escapes the promoter, clearing the promoter and releasing the polymerase to partake in another transcription event. A more detailed mechanistic description is given in Section 2.1.2.

It remains unclear if the steps of the transcription initiation can be modeled as elementary reactions (i.e. having constant probability per time unit to occur, as assumed by the stochastic chemical kinetics) or not. However, if the process is sequential in nature, each of the mechanistic steps themselves must consist of some sequence of elementary steps, making the model appropriate.

When the number of RNA polymerases is approximately constant, the variable R can be eliminated from the model. Such assumption is good when there is a large pool of polymerases available for the transcription, which is a good approximation under favorable growth conditions (Bremer et al., 2003). In this case, it can be shown that the above model is equivalent to the following model in the dynamics of E:

$$P \xrightarrow{k_1} I_1 \xrightarrow{k_2} \cdots I_{n-1} \xrightarrow{k_n} P + E$$

where $I_j \in \mathbb{Z}_1$ are some intermediate complexes of transcription initiation. For strong promoters, the rates k_j are expected to correspond to the forward rates of the McClure (1985) model (Record et al., 1996). Otherwise, the rates of the subsequent steps are affected.

The slow steps in the model are called rate limiting, as they determine the overall rate at which transcription is initiated (McClure, 1985). The fast steps can be neglected, as they play no significant role in determining the dynamics. Meanwhile, the number and relative durations of the rate limiting steps determine the stochasticity in the transcript production independently of the mean: when the transcription rate is limited by a single elementary reaction, the transcription is Poissonian, while a sequence of several rate limiting steps results in more deterministic (i.e. periodic-like) transcription, exhibiting Gaussian-like rather than exponential intervals.

As the above model of transcription initiation consists of a sequence of elementary reactions, the intervals between consecutive transcription initiations are sums of exponential variates. The PDF of the resulting intervals can be obtained by convolving the individual PDFs (Kandhavelu et al., 2011):

$$f(x) = \sum_{i=1}^{n} \left(\prod_{\substack{j=1,\cdots,n \\ j \neq i}} \frac{k_j}{k_j - k_i} \right) k_i \, \exp(-k_i \, x)$$

where $f(x)$ represents the PDF of the transcription initiation intervals, and k_i are the rates of the steps in the model. Note that the order of the steps cannot

be determined without information of the promoter states P, I_1, $\cdots$, I_{n-1}. The mean and variance of the intervals are:

$$\mu = \sum_{i=1}^{n} \frac{1}{k_i}$$
$$\sigma^2 = \sum_{i=1}^{n} \frac{1}{{k_i}^2}$$

which follow from the mean and variance of independent random variables. This always results in a squared coefficient of variation (cv-squared, variance over mean squared) of $\sigma^2 / \mu^2 \leq 1$ (Kandhavelu et al., 2011) with equality if and only if (iff) only one of the k_i is slow. The Poissonian model presented earlier features exponentially distributed inter-transcription intervals, resulting in a cv-squared of unity, regardless of the rate parameter. As the sequential model always results in less noise than a Poissonian production would have (unless the model degenerates the Poissonian production), it is called sub-Poissonian.

The differences to the Poissonian model is demonstrated in Figure 3.1. The density of the distribution resembles Gaussian-like rather than an exponential distribution of the Poissonian model, lacking probability mass from small and large values. Such distribution results in more regular intervals between the transcript production, as demonstrated by the production time traces shown in the lower panels. Production events, which occur in short succession, appear to be absent in the case of sub-Poissonian RNA production, as could be predicted from the fact that the density is zero at $x = 0$, provided that there is more than one rate limiting step.

More recently, measurements in live *E. coli* have presented evidence that such a model can well describe the RNA production of some promoters under specific conditions (Kandhavelu et al., 2011, 2012b,a; Makela et al., 2013; Muthukrishnan et al., 2012). Based on these studies, several of the steps are hypothesized to be rate limiting, allowing temperature, catalysts, and inhibitory molecules to independently control the expression rate and the associated stochasticity of the gene (Kandhavelu et al., 2012a; Muthukrishnan et al., 2012).

3.3.2 Active-inactive promoter model

As the sequential model of transcription always results in sub-Poissonian transcription dynamics, it cannot explain the measurements where highly stochastic RNA numbers have been reported (Chong et al., 2014; So et al., 2011; Taniguchi et al., 2010). Instead, it has been hypothesized that such distributions result from RNAs being produced in bursts (or pulses), which would explain the increased amount of stochasticity.

For such observations, a model where the promoter transits between active and inactive states has been proposed (Kaern et al., 2005; Peccoud and Ycart, 1995;

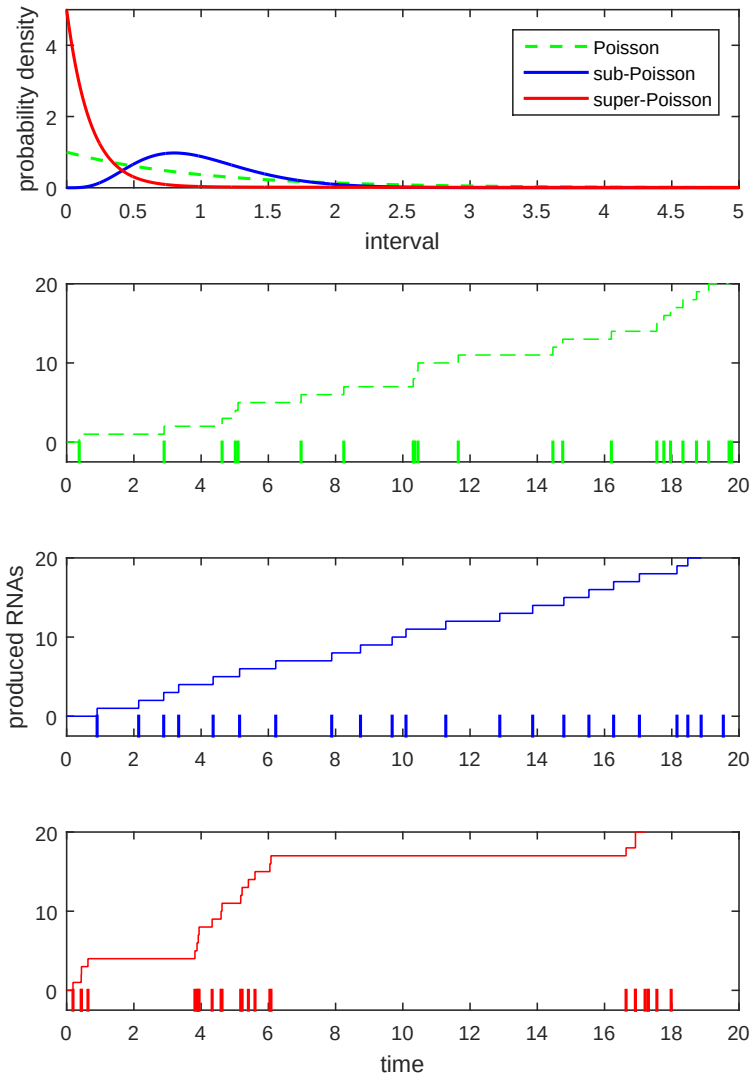

Figure 3.1: Top panel: probability density functions of inter-transcription interval distributions for Poissonian, sub-Poissonian, and super-Poissonian transcription models. Rest of the panels from top to bottom: Number of produced RNAs from MC simulations with Poisson, sub-Poisson, or super-Poissonian transcription model. The colored ticks in the bottom represent jump positions. The expected interval is 1 for each model, and the variance is 1, 1/10, and 10 for the Poissonian, sub-, and super-Poissonian models, respectively.

Shahrezaei and Swain, 2008). Such an active-inactive promoter can be represented by the reactions:

$$P_0 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P \xrightarrow{k_2} P + E$$

where $P_0,\ P \in \mathbb{Z}_1$ represent the inactive and active states of the promoter, respectively, and E is the elongation complex. In such model, the RNA production events are dispersed further from those of the Poissonian model by the gene randomly turning off, resulting in alternating periods of transcriptional activity and inactivity.

There are various hypotheses of what could be the governing mechanism behind such a model. One viable hypothesis is that the transitions between the active and inactive states result from random binding and unbinding of transcription factors (Golding and Cox, 2006; Kaern et al., 2005). A recent study in *E. coli* suggests that such transitions can also result from DNA supercoiling buildup (Chong et al., 2014). Meanwhile, in eukaryotes, it has been hypothesized such transitions are due to eukaryotic chromatin remodeling (Cairns, 2009; Chubb and Liverpool, 2006), or could be due to promoter proximal pausing (Chubb and Liverpool, 2006), which is common for the eukaryotic RNA polymerase II (Wu and Snyder, 2008), or could event result from pauses during elongation (Chubb and Liverpool, 2006). The multitude of hypotheses suggests that there are likely different mechanisms capable of exhibiting such behavior. However, the model is valid regardless of the mechanism responsible for the active-inactive transitions, as long as the transitions occur with constant probability per unit time.

Meanwhile, the transcription intervals resulting from this model have a PDF of:

$$f(x) = \left(\tfrac{k_1 - p^-}{k_1}\tfrac{p^+}{p^+ - p^-}\right)\, p^-\, \exp(-p^-\, x) + \left(\tfrac{p^+ - k_1}{k_1}\tfrac{p^-}{p^+ - p^-}\right)\, p^+\, \exp(-p^+\, x)$$

with

$$p^{\pm} = \frac{k_1 + k_{-1} + k_2}{2} \pm \frac{\sqrt{(\,k_1 + k_{-1} + k_2\,)^2 - 4\,k_1\,k_2}}{2}$$

which is a convex mixture of two exponential distributions with rates $p^-,\ p^+$, such that $0 < p^- < k_1 < p^+$. In this distribution, p^- determines the behavior for short intervals (around $x = 0$), and p^+ the behavior of the tail (large x). The resulting mean and variance of the intervals are (Peccoud and Ycart, 1995):

$$\mu = \left(\frac{k_1}{k_1 + k_{-1}}\right)^{-1} \frac{1}{k_2}$$
$$\sigma^2 = \left(1 + 2\,\frac{k_2}{k_{-1}}\left(1 - \frac{k_1}{k_1 + k_{-1}}\right)^2\right)\,\mu^2$$

From the above, it can be seen that the mean and the noise are controlled by some interesting properties: $k_1\,/\,(k_1 + k_{-1})$ represents the fraction of time the

promoter spends in the active state; $1 / k_2$ would be the average time between transcripts if the gene was constantly active; k_2 / k_{-1} is the burst size, that is, the average number of transcription initiations prior transitioning to the inactive state; and $1 / k_1 + 1 / k_{-1}$ is the burst interval, that is, the average time between the starts of two consecutive bursts. Again, the transcription rate can be limited by different components of the model, such as the burst size or the burst rate.

Meanwhile, such a model always results in cv-squared $\sigma^2 / \mu^2 \geq 1$ of the intervals. The equality can be attained in various degenerate cases, such as if the gene is mostly active ($k_1 \gg k_2,\ k_{-1}$) or if the transitions between active and inactive states are much faster than the transcription rate ($k_1,\ k_{-1} \gg k_2$). Again, Figure 3.1 shows the differences to the Poissonian RNA production model. The active-inactive model features frequent small and large values, which create alternating periods of transcriptional inactivity and bursts of RNA production, which are apparent in the bottom panel of the figure.

Interestingly, while the sub-Poissonian dynamics of the sequential model stem from a series of elementary reactions being wired in series, the super-Poissonian dynamics of the active-inactive promoter model can be seen arising from two elementary reactions being wired in parallel. While either family can feature any rate of production, the space of possible levels stochasticity is partitioned by the two topologies.

3.4 Modeling gene networks

Genes operate in complex circuits, where the interactions are formed by their genetic products acting as factors regulating the expression of other genes. Frequently, genes also regulate their own expression, either directly or via some feedback mechanism (Martinez-Antonio and Collado-Vides, 2006). Gene regulation is typically implemented using proteins (Alberts et al., 2014), but can also function e.g. via small RNAs (Storz and Gottesman, 2006). These genetic interactions result in networks with intricate control structures (Wolf and Arkin, 2003), cooperativity, and various feedback and feedforward connections (Martinez-Antonio and Collado-Vides, 2006). Further, the interactions must be optimized to offer a trade-off between resistance to noise, information loss, and signaling rates (Lestas et al., 2010).

The stochastic modeling strategy of genetic networks (Ribeiro et al., 2006) models the expression of each gene individually, and the binding and unbinding of the transcription factors explicitly. This allows accounting for low copy number effects of the regulatory molecules: the occupancy of the binding site cannot be time averaged, as it can result in bursting of the regulated gene as demonstrated in the previous section (Peccoud and Ycart, 1995), and a single transcription factor can only occupy a single operator site at a time.

Consider a network consisting of n genes. Here, $P_{j,\mathbf{s}_j}$, with $\mathbf{s}_j \doteq \left(s_{1,j}, \cdots, s_{k_j,j} \right)^*$, denotes the state of a promoter region of the gene j with a total of k_j operator sites, where $s_{i,j} = 1$ iff the operator site i is occupied by the appropriate transcription factor, and $s_i = 0$ otherwise.

The expression of each gene is modeled by their appropriate reactions. In addition to the synthesis of RNA and proteins, their degradation should be accounted for, as the products tend not to be long lived (Bernstein et al., 2002; Taniguchi et al., 2010). For example, based on Poissonian transcription and translation, an appropriate model for a single gene could be:

$$\begin{aligned} P_{j,\mathbf{s}_j} + R_j &\xrightarrow{f_j(\mathbf{s}_j)\, k_{M_j}} P_{j,\mathbf{s}_j} + R_j + M_j \quad \text{for } \mathbf{s}_j \in {\mathbb{Z}_1}^{k_j} \\ M_j &\xrightarrow{d_{M_j}} \emptyset \\ M_j + R' &\xrightarrow{k_{X_j}} M_j + R' + X_j \\ X_j &\xrightarrow{d_{X_j}} \emptyset \end{aligned}$$

where M_j and X_j represent the messenger RNA and the protein encoded by the gene j, respectively. Here, R_j is the RNA polymerase holoenzyme of the appropriate type for the gene. The first reaction represents transcription, and it can only occur if the gene is in the appropriate state, as determined by $\mathbf{s}_j$, and an RNA polymerase is available to transcribe it. The transcription rate is determined by the regulatory function $f_j(\mathbf{s}_j)$, which depends on the states of the operator sites, the maximal transcription rate for the gene k_{M_j}, and the RNA polymerase concentration. The second reaction models RNA degradation with an average RNA lifetime of ${d_{M_j}}^{-1}$. Meanwhile, the third reaction represents translation, which is proportional to the abundance of the messenger RNAs and the ribosomes R'. Here, the rate k_{X_j} represents the number of proteins produced per RNA and unit time. Finally, the fourth reaction represents protein degradation, with an average protein lifetime of ${d_{P_j}}^{-1}$.

Alternatively, transcription may take e.g. the form of the sequential model of transcription initiation presented earlier. Such a model can be modeled by semi-Markovian kinetics, which can be efficiently simulated using the delayed SSA (Roussel and Zhu, 2006). Particularly, if there are n steps of approximately equal duration, the transcription intervals are gamma distributed, as opposed to the exponential intervals of the Poissonian model. The kinetic parameters of each gene can be selected to be different, as indicated by the subscripts in the parameters. Also, distinction can be made e.g. between the different kind of RNA polymerase holoenzymes for different sigma factors, which recognize different genes.

The coupling between the different genes is represented by a set of reactions of the form:

$$P_{j,\left(s_{1,j},\cdots,s_{i-1,j},0,s_{i+1,j},\cdots,s_{k_j,j} \right)^*} + m_{i,j}\, X_{r_{i,j}} \underset{b_{i,j}}{\overset{a_{i,j}}{\rightleftharpoons}} P_{j,\left(s_{1,j},\cdots,s_{i-1,j},1,s_{i+1,j},\cdots,s_{k_j,j} \right)^*}$$

which denotes the association and disassociation of the transcription factor to the operator site i of gene j. Here, the transcription factor is an $m_{i,j}$-oligomer of proteins of the species $X_{r_{i,j}}$. It is also possible to form heteromeric oligomers by replacing the protein with an appropriate set of proteins. Meanwhile, $a_{i,j}$ and $b_{i,j}$ represent the association and disassociation rate constants of the transcription factor.

This scheme is readily exemplified e.g. by modeling a n-gene repressilator, which is a genetic circuit where each gene inhibits the expression of the next gene in a ring-form topology. Such circuits are found in cells and are used in keeping track of time, in adapting to periodic environmental conditions (e.g. circadian clock), in controlling information flow e.g. in neurons, and in signal modulation and multiplexing (Elowitz and Leibler, 2000; Wolf and Arkin, 2003). Using the above modeling strategy, a repressilator can be represented by the following reactions, for $j \in \{1, \cdots, n\}$ and $r_j = j-1$ for $j > 1$ and $r_j = n$ for $j = 1$:

$$
\begin{aligned}
P_{j,0} + R_j &\xrightarrow{k_{M_j}} P_{j,0} + R_j + M_j \\
M_j &\xrightarrow{d_{M_j}} \emptyset \\
M_j + R' &\xrightarrow{k_{X_j}} M_j + R' + X_j \\
X_j &\xrightarrow{d_{X_j}} \emptyset \\
P_{j,0} + m_j\, X_{r_j} &\underset{b_j}{\overset{a_j}{\rightleftharpoons}} P_{j,1} \\
P_{j,1} &\xrightarrow{m_j\, d_{X_j}} P_{j,0} + (m_j - 1) X_{r_j}
\end{aligned}
$$

where the first reaction represents transcription, provided that the promoter is not repressed by the repressor X_{r_j}. The third, fourth, and fifth reaction represent RNA degradation, translation, and protein degradation, respectively, as described above. The pair of reactions represents the binding and unbinding of the repressors, and the last reaction models the gene becoming unrepressed due to degradation of one of the proteins in the repressor complex, respectively.

There are ways to reduce the complexity of the above modeling strategy, if some approximations can be made. For example, if the number of the transcription factors is also approximately constant, which is true for abundant transcription factors, the rate at which the binding site becomes occupied is insensitive to the variations in the transcription factor concentration. This implies that the regulation can be modeled using the active-inactive promoter model (see Section 3.3.2), as the transitions are expected to occur with a constant probability per unit time. In some cases, the transcription factor binding can be taken to be much faster than transcription. For example, noise suppressing feedback loops must be sufficiently fast to serve their purpose (Gronlund et al., 2013). In the case where

these assumptions are applicable, the production can be approximated by:

$$R_j \xrightarrow{\overbrace{\frac{k_1}{k_1+k_{-1}}}^{\bar{f}_j}\overbrace{k_2}^{k_{M_j}}} R_j + M_j$$

where the kinetic parameters k_1, k_{-1}, and k_2 are as in Section 3.3.2. With this approximation, f_j is no longer a dynamic property, but has been replaced by the time average $\bar{f}_j$ of the gene being in the active state. Depending on whether the transcription factor regulating the gene is an activator or repressor, here either k_1 or k_{-1} ought to be proportional to the transcription factor concentration. In the case of activation, the fraction of time spent in the active state is given by the Hill function (Hill, 1910):

$$\bar{f}_j = \frac{k_1}{k_1 + k_{-1}} = \frac{X_{r_j}^{m_j}}{X_{r_j}^{m_j} + \underbrace{k_{-1} \,/\, k_1'}_{K^{m_j}}}$$

where $k_1 \doteq k_1' \, X_{r_j}^{m_j}$ is the effective binding rate, and $K \doteq (k_{-1} \,/\, k_1')^{1/m_j}$ is the microscopic disassociation constant, which determines the concentration of X_{r_j} producing half the occupation of the operator site, resulting in half the activity of the gene. In this context, m_j is called the Hill coefficient, and represents the cooperativity of binding. If the transcription factor X_{r_j} was a repressor instead, one would have $k_{-1} \doteq k_{-1}' \, X_{r_j}^{m_j}$, and the activity of the gene is given by the inverse (complementary) Hill function:

$$\bar{f}_j = \frac{k_1}{k_1 + k_{-1}} = \frac{k_1 \,/\, k_{-1}'}{k_1 \,/\, k_{-1}' + X_{r_j}^{m_j}}$$

where similarly, $(k_1 \,/\, k_{-1}')^{1/m_j}$ is the concentration producing half the activity of the gene. If there are multiple independent operator sites for the transcription factors to bind, as opposed to cooperative binding ($m_j > 1$), the fraction is given by the product of the individual functions:

$$\bar{f}_j = \bar{f}_{j,1} \cdots \bar{f}_{j,k_j}$$

Essentially, using the above approximation results in a hybrid model, in which the RNA numbers are modeled using stochastic chemical kinetics, while the protein numbers are modeled with deterministic kinetics. This is a good approximation if the protein numbers exists in sufficiently high concentrations, while the RNA numbers are subject to low-copy number variations. This tends to be true in *E. coli*, where the RNA numbers range from 0.05 to 5 per cell, while protein numbers are 10^2 to 10^4 times higher (Taniguchi et al., 2010). In some cases it might be appropriate to model some of the promoter state transitions using the

stochastic (exact) method, while using the deterministic kinetics for the other state transitions. The advantage of this strategy is that the frequent binding and unbinding of the transcription factors need not to be simulated explicitly.

Finally, an example of this hybrid strategy is given by simulating a model of a three-gene repressilator. The following model is based on the deterministic model presented in Elowitz and Leibler (2000), with equivalent rate constants. The model demonstrates the oscillatory behavior of the synthetic three-gene (LacI, TetR, λ cI) repressilator engineered by Elowitz and Leibler (2000). The model consists of the reactions:

$$\begin{aligned}
\emptyset &\xrightarrow{\frac{\alpha-\alpha_0}{1+(p_j/K)^n}+\alpha_0} m_i \\
m_i &\xrightarrow{1\,/\,\tau_m} \emptyset \\
\emptyset &\xrightarrow{\beta\,/\,\tau_m\, m_i} p_i \\
p_i &\xrightarrow{1\,/\,\tau_p} \emptyset
\end{aligned}$$

where the first reaction represents transcription, α being the maximal transcription rate and α_0 being the basal (or leak) expression rate. Here $\alpha = 0.5\ \mathrm{s}^{-1}$, and $\alpha_0 = 5 \times 10^{-4}\ \mathrm{s}^{-1}$. As the genes repress each other, the transcription rate is modulated by the inverse Hill function, with a Hill coefficient of $n = 2$ and disassociation constant of $K = 40$ monomeric proteins. Meanwhile, the second reaction represents the degradation of the messenger RNA, with an RNA half-life of $\tau_m \ln(2) = 2$ min. The third reaction represents translation, $\beta = 20$ representing the average number of proteins produced per RNA. Finally, the fourth reaction represents protein degradation, with a protein half-life of $\tau_p \ln(2) = 10$ min. One of each of these reactions must exist for $(i, j) \in \{(1, 3), (2, 1), (3, 2)\}$ in order to model each of the three genes in the circuit.

The model was simulated using SSA for a total of 1000 min with the parameters specified above. An example time series of the protein numbers p_i of each gene is shown in Figure 3.2. This example demonstrates the oscillatory behavior of the circuit, which could be deduced from the corresponding rate equations (Elowitz and Leibler, 2000), while exemplifying the stochastic (note the fluctuations in the amplitude and period of the oscillations) yet orderly nature of the above model.

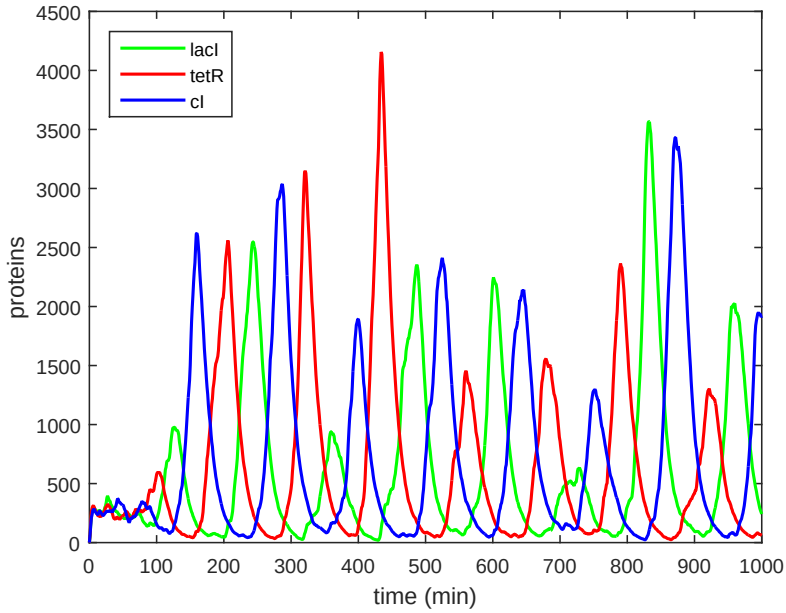

Figure 3.2: Example time series from SSA simulation of the Elowitz and Leibler (2000) repressilator. The three genes, LacI, TetR, and λ cI, each inhibit the expression of the next gene in chain, resulting in oscillations in their protein levels.

4 Statistical methods

4.1 Maximum likelihood estimation

Maximum likelihood (ML) estimation is a parameter estimation technique, which can be applied to arbitrary statistical models. Given a set of data, an ML estimate is a set of parameters, which maximizes the likelihood function. (Lehmann and Casella, 1998; Van Trees et al., 2013) More intuitively, the estimate represents the set of parameters that are most likely to generate the observations. ML estimation was popularized by Fisher between 1912 and 1922, but related techniques were known to earlier authors such as Gauss (Aldrich, 1997). Technically, ML estimation is a parametric technique, implying that a model must be specified; however, the model might be relatively vague, e.g. as it is in the case of the Kaplan-Meier estimator (Kaplan and Meier, 1958; Section 4.4.2).

4.1.1 Estimation theory

Maximum likelihood estimation is not the only way to estimate the parameters of a model, nor is it necessarily the best. Estimation theory (Van Trees et al., 2013) deals with the analysis of estimators, and the most important concepts are revisited here.

In formal terms, $\mathcal{M}(\boldsymbol{\theta})$ is the model, $\boldsymbol{\theta}$ being its parameters. Let $\mathbf{X} \sim \mathcal{M}(\boldsymbol{\theta})$ denote the random variables of interest, and $\mathbf{x}$ is some realization of $\mathbf{X}$. An estimator $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is some function of the data $\mathbf{x}$, which is used to infer the value of the unknown parameters $\boldsymbol{\theta}$ of the model. As the estimator is a function of the data, it is a random variable as well.

Typically, it is preferred that the estimator produces estimates close to the true parameter value. The closeness is measured in terms of the error $\mathbf{e}(\mathbf{x}) \doteq \hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}$ between the estimator and the true value, and some loss function, which is typically taken to be the squared distance.

The expected error $\mathbf{b} \doteq \mathbb{E}[\,\mathbf{e}(\mathbf{X})\,]$ is called bias, and it indicates how much the estimator is off the true value on average (consistently underestimates or overestimates the parameters). An attractive property of an estimator is unbiasedness, which implies that, on average, the estimator correctly estimates the true parameter

value. Such an estimator might or might not exist. Some related properties are asymptotic unbiasedness, which implies that as the sample size increases, the bias converges to zero, and consistency, which implies that the estimator further converges in probability to the true parameter value.

As the estimator is a random variable, it produces different estimates for different realizations of $\mathbf{X}$. Another important property is the estimator (co)variance $\mathbf{c} \doteq \mathbb{C}\mathrm{ov}[\,\hat{\boldsymbol{\theta}}(\mathbf{X})\,]$. The covariance quantifies how dispersed estimates the estimator produces, in terms of squared distance.

The performance, in terms of the squared distance, of any estimator is limited by the Cramér-Rao lower bound (CRLB). The CRLB specifies an information theoretical lower bound on the estimator covariance in terms of its bias:

$$\mathbb{C}\mathrm{ov}[\,\hat{\boldsymbol{\theta}}(\mathbf{X})\,] \geq (\,\mathbf{I} + \mathbf{b}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\,)\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\,(\,\mathbf{I} + \mathbf{b}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\,)^*$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{b}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})$ is the Jacobian matrix of the bias $\mathbf{b}(\boldsymbol{\theta})$, and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is Fisher information matrix. The inequality $\mathbf{A} \geq \mathbf{B}$ is taken to mean that $\mathbf{A}{-}\mathbf{B}$ is positive semidefinite. The bound relies on some mild regularity conditions, for example, the Fisher information must be well defined. An estimator, which is the best possible in some terms is called efficient, typically referring to an estimator attaining the equality in the CRLB. The property of efficiency can also exist either for a finite sample or in asymptotic sense, and there is no guarantee that an efficient estimator exists.

In the above, the Fisher information is:

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}\Big[\,\Big(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}\,|\,\boldsymbol{\theta})\Big)\,\Big(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}\,|\,\boldsymbol{\theta})\Big)^*\,\Big]$$

and it quantifies how much, on average, a sample of the distribution informs about its parameter. The Fisher information plays a key role in the asymptotic theory of maximum likelihood estimation, as will be shown later. In the case of an unbiased estimator, the estimator variance is bounded from below by exactly the inverse of the Fisher information.

When designing the estimator, it is possible to trade bias for variance. For example, one could devise a crude estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{0}$, which would have zero variance, but likely a large bias. Both the bias and the estimator variance are captured by the mean squared error $\mathbb{E}[\,\mathbf{e}(\mathbf{X})^*\,\mathbf{e}(\mathbf{X})\,] = \mathbf{b}^*\,\mathbf{b} + \mathbf{c}$, which is the average (squared) distance between the estimates and the true value. Here, the first term penalizes being regularly off the true value (squared norm of the bias) and the second term the dispersion of the estimates (estimator variance). Again, for an unbiased estimator, this statistic is solely captured by the estimator variance.

4.1.2 ML estimation

A maximum likelihood estimator is defined to be a set of parameters, which maximize the likelihood function. Formally, let $f(\mathbf{x}\,|\,\boldsymbol{\theta})$ be the joint probability

density function (PDF) of the data $\mathbf{x}$ for some set of parameters $\boldsymbol{\theta}$. When this density is considered from the perspective of varying $\boldsymbol{\theta}$, it is called likelihood. Typically, it is more comfortable to work with some derived quantity, such as, the average log-likelihood:

$$\bar{\ell} = \frac{1}{n} \log f(\mathbf{x} \,|\, \boldsymbol{\theta})$$

where n is an appropriate normalizing factor. This is convenient when the PDF can be factored, or is of exponential form. For example, if the data are independent and identically distributed (iid), the joint density can be factored into a product of the marginal densities, and it simplifies to:

$$\bar{\ell} = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \,|\, \boldsymbol{\theta})$$

Now, an ML estimator $\hat{\boldsymbol{\theta}}_{\text{mle}}$ is a set of parameters $\boldsymbol{\theta}$ which maximize the likelihood $f(\mathbf{x} \,|\, \boldsymbol{\theta})$, that is:

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} \; f(\mathbf{x} \,|\, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \; \bar{\ell}$$

where the latter equality follows from the fact that the maxima of f and $g(f)$ for strictly monotonic increasing g are located at the same points. Such maxima must occur where the gradient of $\bar{\ell}$ is zero and the Hessian of $\bar{\ell}$ is non-positive, or at the boundaries of the parameter space. In general, ML estimator need not to be unique, but for many practical problems it is.

Under some mild regularity conditions, such as provided that the model is identifiable, ML estimators feature several attractive properties. These properties include consistency and efficiency. Also, with some further regularity conditions, an ML estimator is asymptotically normal. When all of these properties apply, it follows that at the infinite-sample limit the estimator is normally distributed with a mean equal to the parameter value being estimated and covariance equal to the inverse of the Fisher information matrix, that is:

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1})$$

where $\cdot \xrightarrow{d} \cdot$ denotes convergence in distribution, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a normal distribution with a mean of $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is the Fisher information matrix. This result reveals the link between ML estimation and Fisher information. Regardless of being an asymptotic result, it is often exploited to estimate the uncertainty of the estimated parameters for finite samples, for example, for the purposes of deriving confidence intervals.

Further, due to continuous mapping theorem (Lehmann and Casella, 1998; Mann and Wald, 1943), it can be shown that:

$$\sqrt{n} \left(\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}) \right) \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \, \mathbf{g}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}) \, \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1} \, \mathbf{g}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})^* \right)$$

for any $\mathbf{g}(\boldsymbol{\theta})$ which is continuous almost everywhere. Here, $\mathbf{g}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})$ is the Jacobian of $\mathbf{g}(\boldsymbol{\theta})$. This form is useful in estimating the uncertainty of arbitrary model features after ML estimation.

Finally, ML estimators are invariant on some transformations. These include the transformation on the parameters, that is, if $\hat{\boldsymbol{\theta}}$ is the ML estimator for $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\alpha}} = \mathbf{g}(\hat{\boldsymbol{\theta}})$ is the ML estimator for $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, which is due to the chain rule of differentiation. Meanwhile, one-to-one transformations $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ on the data result in a PDF of $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{y}) \,/\, |\det(\mathbf{g}_{\mathbf{x}^*}(\mathbf{y}))|$, where $\mathbf{g}_{\mathbf{x}^*}(\mathbf{y})$ is the Jacobian of $\mathbf{g}(\mathbf{x})$ at $\mathbf{y}$. If $\mathbf{g}(\mathbf{x})$ is constant with respect to $\boldsymbol{\theta}$, so is its Jacobian, and the estimator is unchanged by the transformation, provided that the appropriate transformation on the model is applied.

4.1.3 ML estimation for the exponential family

Many common densities, such as (multivariate) normal distribution, gamma distribution, and Poisson distribution belong to the exponential family. Technically this means that their PDF can be decomposed as follows:

$$f(\mathbf{x} \,|\, \boldsymbol{\theta}) = b(\mathbf{x}) \, \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^* \, \mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta}(\boldsymbol{\theta})))$$

where $b(\mathbf{x})$ is called the base measure, $\boldsymbol{\eta}$ the natural parameters, $\mathbf{t}(\mathbf{x})$ the sufficient statistic, and $A(\boldsymbol{\eta})$ the log-partition function. (Lehmann and Casella, 1998) The sufficient statistic $t_i(\mathbf{x})$ summarizes all the information on the data $\mathbf{x}$ about the parameter η_i, while $b(\mathbf{x})$ and $A(\boldsymbol{\eta})$ are just normalizing factors in the sample and parameter spaces, respectively. Many closed form solutions can be found for variables with such PDFs. If the dimension of $\boldsymbol{\eta}$ is equal to (less than) that of $\boldsymbol{\theta}$, the distribution is said to be of regular (curved) type. The distributions of the curved type lack many of the attractive properties.

An important feature of this type of distributions is that the derivatives of A and the moments of the sufficient statistics are related via:

$$\begin{aligned} \log \mathbb{E}[\, \exp(\mathbf{s}^* \, \mathbf{t}(\mathbf{X})) \,] &= A(\mathbf{s} + \boldsymbol{\eta}) - A(\boldsymbol{\eta}) \\ &= \frac{1}{1!} \, \mathbf{s}^* \, \underbrace{A_{\boldsymbol{\eta}}(\boldsymbol{\eta})}_{\mathbb{E}[\, \mathbf{t}(\mathbf{X}) \,]} + \frac{1}{2!} \, \mathbf{s}^* \, \underbrace{A_{\boldsymbol{\eta}\boldsymbol{\eta}^*}(\boldsymbol{\eta})}_{\mathbb{C}\mathrm{ov}[\, \mathbf{t}(\mathbf{X}) \,]} \, \mathbf{s} + \cdots \end{aligned}$$

which is the cumulant generating function of $\mathbf{t}(\mathbf{x})$. Here $A_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ and $A_{\boldsymbol{\eta}\boldsymbol{\eta}^*}(\boldsymbol{\eta})$ represent the gradient and the Hessian of $A(\boldsymbol{\eta})$, respectively.

The average log-likelihood for n iid samples is:

$$\bar{\ell} = \frac{1}{n} \sum_{i=1}^{n} \log \, b(\mathbf{x}_i) + \boldsymbol{\eta}^* \, \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbf{t}(\mathbf{x}_i)}_{\bar{\mathbf{t}}} - A(\boldsymbol{\eta})$$

which demonstrates that the average of the sufficient statistic $\bar{\mathbf{t}}$ (or equivalently the sum) can summarize arbitrary amounts of data $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ using only a fixed number of values. The resulting ML estimator is the root of $\bar{\mathbf{t}} - A_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \bar{\mathbf{t}} - \mathbb{E}[\,\mathbf{t}(\mathbf{X})\,]$. Interestingly, such ML estimator equates the average sufficient statistics of the data with that of the expectation from the model. The Hessian of the ML objective is $\bar{\ell}_{\boldsymbol{\eta}\boldsymbol{\eta}^*} = -A_{\boldsymbol{\eta}\boldsymbol{\eta}^*}(\boldsymbol{\eta}) = -\,\mathbb{C}\mathrm{ov}[\,\mathbf{t}(\mathbf{X})\,]$, which implies that the objective $\bar{\ell}$ is concave, and the maximum is unique, unless the distribution is degenerate (that is, of the curved type).

The following serves as a simple example of ML estimation. Suppose an iid exponentially distributed sample $x_1, \cdots, x_n \sim \mathcal{E}(\lambda)$ with a rate of λ, which is to be estimated. Since the PDF for exponential distribution is $f(x \,|\, \lambda) = \lambda \exp(-\lambda x)$, the average log-likelihood is:

$$\bar{\ell} = -\lambda \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)}_{m} + \log \lambda$$

where m is the sample mean. The first derivative of $\bar{\ell}$ is $\bar{\ell}_\lambda = -m + \lambda^{-1}$ and the second derivative is $\bar{\ell}_{\lambda\lambda} = -\lambda^{-2}$, implying that the ML estimator is $\hat{\lambda} = m^{-1}$. The asymptotic properties imply that $\hat{\lambda}^{-1}$ converges in distribution to $\mathcal{N}(\lambda^{-1}, \lambda^{-2} \,/\, n)$ and $\hat{\lambda}$ to $\mathcal{N}(\lambda, \lambda^2 \,/\, n)$ as $n \to \infty$. Clearly, the estimator $\hat{\lambda}$ is consistent and asymptotically normal, and, as the Fisher information is $n \,/\, \lambda^2$, it also attains the CRLB asymptotically, making it efficient. However, these properties only hold asymptotically, and for example, for $n = 1$ the estimator has infinite bias, variance, and higher-order moments.

Note that the exponential distribution is also of regular exponential family. Here, the natural parameter is $\eta = -\lambda$, the sufficient statistic $t(\mathbf{x}) = m$ is the sample mean, and the normalization factors are $b(\mathbf{x}) = 1$ and $A(\eta) = -\log(-\eta)$. As $A_\eta(\eta) = -\eta^{-1} = \lambda^{-1}$, a choice of parameters which results in an unbiased estimator attaining the CRLB would be $\hat{\lambda}^{-1} = m$. This estimator is unbiased and of minimum variance for all sample sizes, including $n = 1$, but it is normal only in the asymptotic sense.

The properties derived above can be further exemplified using Monte Carlo (MC) simulations. In the simulations, 1, 10, and 100 samples were generated from an exponential distribution with a mean of 1. The distributions of the estimates both for the mean λ^{-1} and the rate λ of the exponential distribution were obtained using the above ML estimators. These distributions are shown in Figure 4.1. The statistics were gathered from 10^6 simulations.

In case of 100 samples, the estimates are well approximated by the normal distribution that follows from the asymptotic ML theory: the estimates for rate and mean have a mean (variance) of 1.01 (0.01) and 1.00 (0.01), respectively. For 10 samples, the distributions of estimates are skewed, and consequently clearly not

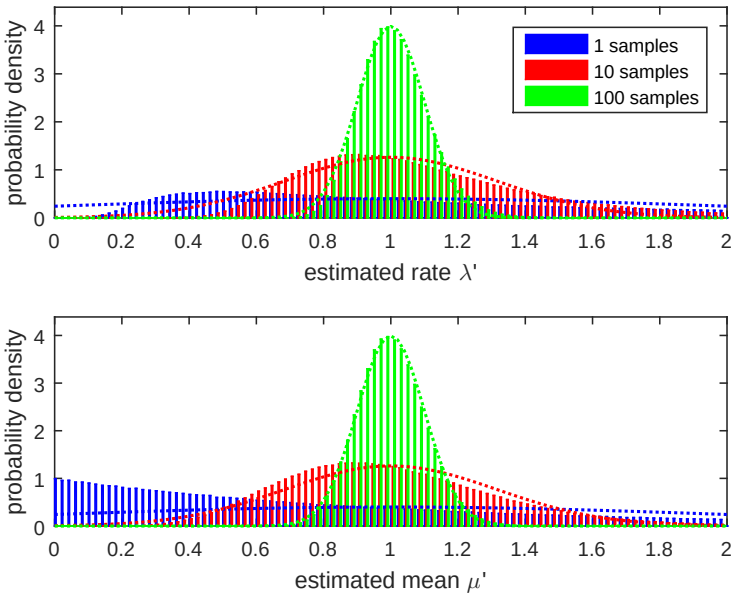

Figure 4.1: Upper panel: Histograms of estimates (bars) of the rate parameter from 1, 10, and 100 samples. Lower panel: Histograms of estimates (bars) of the mean parameter from 1, 10, and 100 samples. The dashed lines represent the expected distributions, which result from the asymptotic properties of the ML estimator.

normal, and a bias exists in the rate estimates. In this case, the means (variances) are 1.11 (0.15) and 1.00 (0.10) for rate and mean estimates, respectively. These effects are corroborated when using just 1 sample. The mean (variance) from the MC simulations is 15.73 (7.00×10^6) for the rate estimates, and 1.00 (1.00) for the estimates. The former estimator is a poor estimator, as the estimates are expected to be infinitely far from the true value. As such, its mean and variance cannot be reliably estimated, and the above numbers have little meaning. Meanwhile, in the latter case, the estimates are exponentially distributed as the estimator is the sample itself. More importantly, this estimator is unbiased and of minimum variance, and as such, it is the best possible estimator in terms of squared distance to the true value that one can devise.

4.2 The expectation maximization algorithm

The expectation maximization (EM) algorithm is an iterative method for ML estimation. The method was introduced by Dempster et al. (1977) in its general form, but similar techniques have been used earlier for more specific problems (e.g. Sundberg, 1974). The convergence properties of the general EM algorithm were established by Wu (1983).

The algorithm can be applied when the marginal density of the model is intractable for a direct solution of the ML problem, but a simpler model can be obtained by introducing latent variables. This happens to be true for many statistical problems, which arise from practical situations. A typical scenario is when the data come from mixtures of simple distributions. This could result, for example, if the measurements come from distinct populations, each of which can be modeled by a relatively simple model, and it is not known to which population the samples belong to.

The idea of the algorithm is to find the ML estimate by starting with some initial estimate of the parameters, and to seek for a sequence of increments in the likelihood function. Such increments are obtained by estimating the latent parameters with the aid of the current parameter estimate, followed by finding improved parameters for the observed data and the estimated latent variables.

4.2.1 Generalized EM algorithm

Suppose that there is some complete set of data $(\mathbf{x}, \mathbf{z})$, where $\mathbf{x}$ are the observed data, and $\mathbf{z}$ are the latent data. The observed data $\mathbf{x}$ are called "incomplete data", since they contain a subset of the information of the "complete data" $(\mathbf{x}, \mathbf{z})$. The model for the complete data is specified by the joint density $f(\mathbf{x}, \mathbf{z} \,|\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are its parameters. The marginal density for the observed data $\mathbf{x}$ is:

$$g(\mathbf{x} \,|\, \boldsymbol{\theta}) = \int_{\Omega_{\mathbf{Z}}} f(\mathbf{x}, \mathbf{z} \,|\, \boldsymbol{\theta}) \; \partial \mathbf{z}$$

where $\Omega_{\mathbf{Z}}$ is the sample space of $\mathbf{Z}$, the random variable associated with the latent data $\mathbf{z}$. Here, the objective is to estimate the model parameters $\boldsymbol{\theta}$ using the ML estimator:

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} \log \, g(\mathbf{x} \,|\, \boldsymbol{\theta})$$

as the latent data $\mathbf{z}$ are not known. However, in many cases a direct solution might not be easy to find, since the marginal density g might be a combination of many different models. However, if the complete data model f itself admits an easy form, one might be able to make progress.

The EM algorithm is based on the following expansion:

$$\log \, g(\mathbf{x} \,|\, \boldsymbol{\theta}) = \underbrace{\mathbb{E}\Big[\log f(\mathbf{x}, \mathbf{Z} \,|\, \boldsymbol{\theta}) \,\Big|\, \mathbf{x}, \boldsymbol{\theta}^{(p)} \Big]}_{Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(p)})} + \underbrace{\mathbb{E}\Big[-\log h(\mathbf{Z} \,|\, \mathbf{x}, \boldsymbol{\theta}) \,\Big|\, \mathbf{x}, \boldsymbol{\theta}^{(p)} \Big]}_{H(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(p)})}$$

where $h(\mathbf{z} \,|\, \mathbf{x}, \boldsymbol{\theta})$ is the density of $\mathbf{Z} \,|\, \mathbf{x}$. Here, $Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(p)})$ is the average joint likelihood, given the observed data $\mathbf{x}$ and some parameter estimate $\boldsymbol{\theta}^{(p)}$, and $H(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(p)})$ is the cross entropy of $\mathbf{Z} \,|\, \mathbf{x}, \boldsymbol{\theta}^{(p)}$ with $\mathbf{Z} \,|\, \mathbf{x}, \boldsymbol{\theta}$. As the cross-entropy is bounded from below by the entropy $H(\boldsymbol{\theta}^{(p)} \,|\, \boldsymbol{\theta}^{(p)})$ of $\mathbf{Z} \,|\, \mathbf{x}, \boldsymbol{\theta}^{(p)}$, a change $Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(p)}) - Q(\boldsymbol{\theta}^{(p)} \,|\, \boldsymbol{\theta}^{(p)})$

induces an equally large or a larger change in $\log g(\mathbf{x} \mid \boldsymbol{\theta}) - \log g(\mathbf{x} \mid \boldsymbol{\theta}^{(p)})$. Consequently, given some parameter estimate $\boldsymbol{\theta}^{(p)}$, it is possible to find a new estimate $\boldsymbol{\theta}$ which is no worse in terms of the marginal likelihood by improving on $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$.

The EM algorithm starts with an initial estimate $\boldsymbol{\theta}^{(0)}$, and is typically presented as an iteration of the following two steps: the expectation step (or E-step):

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)}) \doteq \mathbb{E}\Big[\log f(\mathbf{x}, \mathbf{Z} \mid \boldsymbol{\theta}) \,\Big|\, \mathbf{x}, \boldsymbol{\theta}^{(p)} \Big]$$

and the maximization step (M-step):

$$\boldsymbol{\theta}^{(p+1)} \doteq \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$$

where $\boldsymbol{\theta}^{(p)}$ is the parameter estimate after p steps. This iteration results in a sequence of parameter estimates $\boldsymbol{\theta}^{(p)}$ with non-decreasing likelihood, as shown above. Often the two steps are not separate computational steps, but conceptualize the idea of the EM algorithm. Conceptually, maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$ involves finding the latent variables (or, in fact, their distribution, $h(\mathbf{z} \mid \boldsymbol{\theta}^{(p)})$) under the current set of parameters, and then finding an improved parameter estimate using the complete data $(\mathbf{x}, \mathbf{z})$. As any EM sequence increases the likelihood, the procedure must converge provided that the likelihood is bounded from above.

The above outlines the traditional EM algorithm. However, some variants use a point estimate of the latent data $\mathbf{z}$, for example, by selecting $\mathbf{z}$ in ML sense. A well known example of such algorithm is the k-means clustering algorithm (Lloyd, 1957), in which a point is assigned to the cluster with nearest cluster centroid, after which the cluster centroid is updated using the current set of points belonging into it. Such variant is referred to as "hard" EM algorithm, and the traditional one as "soft", due to the fact that they use hard (fixed) or soft (fuzzy) estimate of $\mathbf{Z}$. The hard EM variants are primarily useful due to their lower computational demands.

As stated above, it is not necessary to maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$ to make progress, but any improvement is sufficient. Such a variant is commonly referred to as generalized EM (GEM) algorithm. Relaxing the maximization problem might allow easier derivation of the update scheme, but can have a negative impact on the convergence of the algorithm.

The main advantage of the EM algorithm is that the implementation of the E- and M-steps tend to be easy for many practical problems. This allows a solution to the EM steps to be found in closed form, despite the fact that the direct solution of the ML problem is intractable. There are well known families of problems where this holds, such as when the complete data density f is of an exponential family (Sundberg, 1974). Another advantage is that general numerical methods for solving the direct ML problem, such as gradient descent, Newton's method, or any variations of the two, require evaluations of gradients and/or the Hessians

of the objective function. Meanwhile, any derivatives are not needed for the EM algorithm. Also, EM algorithms are often conservative in using computer memory, which is important for large problems.

One disadvantage is that with multimodal problems the EM algorithm tends to converge to local maxima. It is also possible that the algorithm converges to stationary points, which are not even local maxima, or that it gets trapped due to discontinuities in the objective function (Wu, 1983). These effects can be mitigated by using random restarts and/or stochastic optimization methods such as simulated annealing. Another disadvantage of the method is that, in the general case, there are no bounds for the rate of convergence, and typically the convergence rate is relatively slow for high-dimensional problems, when compared e.g. with Newton's method.

4.2.2 EM algorithm for the exponential family

The EM algorithm for a density f of an exponential family can be implemented as follows (Dempster et al., 1977). Note that, g is generally not of an exponential family, or the ML estimate would be readily available. Recall that the density of an exponential family is:

$$f(\mathbf{x}, \mathbf{z} \,|\, \boldsymbol{\eta}) = b(\mathbf{x}, \mathbf{z}) \, \exp(\boldsymbol{\eta}^* \, \mathbf{t}(\mathbf{x}, \mathbf{z}) - A(\boldsymbol{\eta}))$$

Now, the expected likelihood is:

$$Q(\boldsymbol{\eta} \,|\, \boldsymbol{\eta}^{(p)}) = \mathbb{E}\Big[\log b(\mathbf{x}, \mathbf{Z}) \,\Big|\, \mathbf{x}, \boldsymbol{\eta}^{(p)} \Big] + \boldsymbol{\eta}^* \; \mathbb{E}\Big[\mathbf{t}(\mathbf{x}, \mathbf{Z}) \,\Big|\, \mathbf{x}, \boldsymbol{\eta}^{(p)} \Big] - A(\boldsymbol{\eta})$$

which has an unique maximum with respect to $\boldsymbol{\eta}$ when $A_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \mathbb{E}\Big[\mathbf{t}(\mathbf{x}, \mathbf{Z}) \,\Big|\, \mathbf{x}, \boldsymbol{\eta}^{(p)} \Big]$, provided that f is of regular exponential family. Consequently, maximizing $\log g(\mathbf{x} \,|\, \boldsymbol{\eta})$ results in the following steps. The E-step:

$$\bar{\mathbf{t}}^{(p)} = \mathbb{E}\Big[\mathbf{t}(\mathbf{x}, \mathbf{Z}) \,|\, \mathbf{x}, \boldsymbol{\eta}^{(p)} \Big]$$

and the M-step:

$$\text{find } \boldsymbol{\eta}^{(p+1)} \text{ such that } \; \mathbb{E}\Big[\mathbf{t}(\mathbf{X}, \mathbf{Z}) \,\Big|\, \boldsymbol{\eta}^{(p+1)} \Big] = \bar{\mathbf{t}}^{(p)}$$

which shows that the iteration involves estimating the sufficient statistic $\bar{\mathbf{t}}^{(p)}$ and finding the corresponding parameter $\boldsymbol{\eta}^{(p+1)}$. This is an example of a special case of the EM algorithm, which was known prior to Dempster et al. (1977), and has been discussed e.g. in Sundberg (1974).

The above algorithm can be exemplified with the following problem. Consider n iid samples $\mathbf{x}$ from a mixture of two exponential distributions with means μ_1 and μ_2, and mixing weights of w_1 and $(1-w_1)$. For the problem to be identifiable, it

is required that $0 < w_1 < 1$, $\mu_1, \mu_2 > 0$, and $\mu_1 \neq \mu_2$ (the special cases could be analyzed separately). Let the latent variables $Z_i \in \{1, 2\}$ be the classes of each sample, that is $z_i = k$ if and only if (iff) the ith sample is of the population with the kth distribution. Here, with the lack of any further knowledge, the objective is to estimate the parameters $\boldsymbol{\theta} = (\,\mu_1, \mu_2, w_1\,)^*$. The density of the complete data $(\mathbf{x}, \mathbf{z})$ is:

$$
\begin{aligned}
f(\mathbf{x}, \mathbf{z}) &= \prod_{i=1}^{n} \left(\delta\{\, z_i = 1 \,\} \underbrace{\underbrace{w_1}_{h(z_i=1)} \frac{\exp\!\left(-\frac{x_i}{\mu_1}\right)}{\mu_1}}_{f(x_i, z_i=1)} + \delta\{\, z_i = 2 \,\} (1-w_1) \frac{\exp\!\left(-\frac{x_i}{\mu_2}\right)}{\mu_2} \right) \\
&= \underbrace{1}_{b(\mathbf{x},\mathbf{z})} \exp\left(\underbrace{\begin{pmatrix} -1/\mu_1 \\ -1/\mu_2 \\ \log\!\left(\frac{w_1}{\mu_1} \frac{\mu_2}{1-w_1}\right) \end{pmatrix}}_{\boldsymbol{\eta}} {}^* \underbrace{\begin{pmatrix} \sum_{i=1}^{n} \delta\{\, z_i = 1 \,\} x_i \\ \sum_{i=1}^{n} \delta\{\, z_i = 2 \,\} x_i \\ \sum_{i=1}^{n} \delta\{\, z_i = 1 \,\} \end{pmatrix}}_{\mathbf{t}(\mathbf{x},\mathbf{z})} - \underbrace{n \log\!\left(\frac{1-w_1}{\mu_2}\right)}_{A(\boldsymbol{\eta})} \right)
\end{aligned}
$$

which is clearly of exponential family, while the marginal density $g(\mathbf{x})$ is not. Here, $\delta\{\cdot\}$ is the indicator function. Since the mapping $\boldsymbol{\eta}(\boldsymbol{\theta})$ is one-to-one, there is no need to work with the natural parameter $\boldsymbol{\eta}$; it is possible to analyze the sufficient statistic $\mathbf{t}(\mathbf{x}, \mathbf{z})$ in terms of $\boldsymbol{\theta}$. Here, the E-step is:

$$
\mathbf{t}^{(p)} = \mathbb{E}\Big[\, \mathbf{t}(\mathbf{x}, \mathbf{Z}) \,\Big|\, \mathbf{x}, \boldsymbol{\theta}^{(p)} \,\Big] = \begin{pmatrix} \sum_{i=1}^{n} v_{i,1}^{(p)}\, x_i \\ \sum_{i=1}^{n} \left(1 - v_{i,1}^{(p)}\right)\, x_i \\ \sum_{i=1}^{n} v_{i,1}^{(p)} \end{pmatrix}
$$

$$
\text{where } v_{i,k}^{(p)} \doteq \frac{w_k^{(p)}\, \exp\!\left(-\frac{x_k}{\mu_k^{(p)}}\right) / \mu_k^{(p)}}{w_1^{(p)}\, \exp\!\left(-\frac{x_1}{\mu_1^{(p)}}\right) / \mu_1^{(p)} + \left(1 - w_1^{(p)}\right) \exp\!\left(-\frac{x_i}{\mu_2^{(p)}}\right) / \mu_2^{(p)}}
$$

and $w_2^{(p)} = 1 - w_1^{(p)}$. This step is a simple forward computation of a weighted average. The role of $v_{i,k}^{(p)}$ is the class membership probability for the sample x_i being of the class k at the pth step. Meanwhile, the M-step is:

$$
\mathbb{E}\Big[\, \mathbf{t}(\mathbf{X}, \mathbf{Z}) \,\Big|\, \boldsymbol{\theta}^{(p+1)} \,\Big] = \begin{pmatrix} n\, w_1^{(p+1)}\, \mu_1^{(p+1)} \\ n\, \left(1 - w_1^{(p+1)}\right)\, \mu_2^{(p+1)} \\ n\, w_1^{(p+1)} \end{pmatrix} = \mathbf{t}^{(p)}
$$

Combining the two steps and solving for $\boldsymbol{\theta}^{(p+1)}$ gives the EM iteration of:

$$w_1^{(p+1)} = \frac{1}{n} \sum_{i=1}^{n} v_{i,1}^{(p)}$$
$$\mu_k^{(p+1)} = \left(\sum_{i=1}^{n} v_{i,k}^{(p)} \, x_i \right) / \sum_{i=1} v_{i,k}^{(p)}$$

which is an intuitive solution: the mixing weight is an average of the class membership probabilities, while the means of the two distributions are averages of the samples weighted with the class membership probabilities.

4.3 Likelihood ratio test

Likelihood ratio (LR) test is a statistical test to compare goodness of fit of two models which are nested, that is, the simpler model is a special case of the more complex model. Such comparison requires statistical hypothesis testing, as the more complex model never fits worse. The simpler model is called the null model, as it is the model under which the null hypothesis H_0 is constructed. Meanwhile, the more complex model is involved in the alternative hypothesis H_1. (Van Trees et al., 2013) Hypothesis testing allows the null model to be rejected, if the complex model fits much better in a statistically significant sense; otherwise, either the null model is the more appropriate model, or the data lacks statistical evidence to reject the null model. Many common statistical tests such as the Z-test, the F-test, and Pearson's chi-squared test are in fact LR tests with particular null and alternative models.

In the LR test, the likelihood ratio is defined as:

$$\Lambda \doteq \frac{L_0}{L_1}$$

where L_0 and L_1 are the likelihoods of the null and the alternative model. The null hypothesis is rejected if the likelihood ratio is small: $\Lambda < c$. The threshold c is selected based on the null model, and some significance level $\alpha \doteq \mathbb{P}[\, \Lambda < c \,|\, H_0 \,]$, the probability of falsely rejecting the null hypothesis.

If the hypotheses are composite, that is, the model parameters are not fixed, but are contained in some parameter spaces $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ for the null and alternative hypotheses, respectively, the likelihood ratio is computed as follows:

$$\Lambda \doteq \frac{\sup \{\, L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \,\}}{\sup \{\, L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 \,\}}$$

where $\sup \{\cdot\}$ denotes the supremum, that is, the least value not less any of those in the set. Consequently, a likelihood ratio Λ determined by the likelihoods of

two models at their ML estimates corresponds to a hypothesis test between the corresponding families of models.

Such tests are attractive, since the Neyman-Pearson lemma states that they are the most powerful statistical test (Neyman and Pearson, 1933) at a particular significance level. Here, power means statistical power, that is, the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true, $\beta \doteq \mathbb{P}[\,\Lambda < c \,|\, H_1\,]$. The lemma applies for simple hypotheses or in particular cases of composite hypotheses.

Typically, it is difficult to find the distribution of the likelihood ratio Λ for arbitrary models. Therefore, either approximations or numerical methods, such as Monte Carlo simulations need to be used. One frequently used approximation is based on Wilks' theorem, which states that, regardless of the models being compared, a particular statistic:

$$D = -2 \log \Lambda = -2n\,\bar{\ell}_0 + 2n\,\bar{\ell}_1$$

follows a chi-squared distribution with $d_1 - d_0$ degrees of freedom in the infinite sample limit $n \to \infty$ (Wilks, 1938). Here d_0 and d_1 are the number of parameters in the null and the alternative model, respectively. This asymptotic distribution can be used to compute an empirical p-value, which can be used to determine if the null hypothesis can be rejected. As the approximation is good for large sample sizes regardless of the models, this asymptotic result is often used to compute p-values in practice.

4.4 Survival analysis

Survival analysis deals with the analysis of events where part of the population survives past a certain time and the other part does not (Lawless, 2003). The objective is to analyze the changes in the fractions of the two populations, to determine which kind of traits can increase or decrease the probability of survival, or how long is the surviving population still expected to survive. Such analysis is used in e.g. medical testing, where it is not feasible to wait until all patients have died, and reliability analysis, for similar reasons (Lawless, 2003). Consequently, the events are called deaths or failures as they relate to the death of a patient or the failure of a machine in the two fields, respectively. However, in the context of this thesis "survival" would correspond to certain event not being observed in the measurement time window.

4.4.1 Truncation and censoring

Truncation is a condition where values of a random variable are only observed if they occur in some region of the sample space. For example, if X is some random variable, its truncation by B is $Y \doteq X \,|\, (X \in B)$, where B is some subset of the sample space of X. Truncated values can be generated e.g. using rejection sampling (see Section 3.2.2).

If small (large) values are not observed, the process is called left (right) truncation, as the values are bounded from left (right). If both small and large values are not observed, the values are said to be interval truncated, as the values are limited to an interval. For example, a common case for left truncation to occur is when the gathering of data begins retrospectively to the experiment.

The probability of observing a truncated value is:

$$\mathbb{P}[\,Y \simeq x\,] = \mathbb{P}[\,X \simeq x \,|\, X \in B\,] = \begin{cases} \mathbb{P}[\,X \simeq x\,]\ \mathbb{P}[\,X \in B\,]^{-1} & \text{, for } X \in B \\ 0 & \text{, otherwise} \end{cases}$$

where $\mathbb{P}[\,X \simeq x\,] \simeq \mathbb{P}[\,x \leq X < x + \partial x\,]$ is the probability that X is in some infinitesimal region around x, given by $f_X(x)\ \partial x$ if f_X is the PDF of X.

A similar, but more informative condition is called censoring. In censoring, instead of the value being completely unobserved, incomplete information of the event is observed. Typically, the values which satisfy some condition are exactly observed, and for others, it is apparent that the condition is not true. Again, the conditions where the value is limited from below, above, or from both sides are called left, right censoring, and interval censoring, respectively.

Formally, let $Z = \phi(Y)$, where the function ϕ censors Y by partitioning B. Let $A = \phi^{-1}(\{Z\}) = \{\, y \in B : \phi(y) = \phi(Y)\,\}$ denote the preimage of $\{Z\}$. Here, Y is said to be censored into A, as it is the set of possible values of Y after observing Z. The probability is given by:

$$\mathbb{P}[\,Z \simeq z\,] = \mathbb{P}[\,\phi(Y) \simeq z\,] = \sum_{\Omega \subseteq \phi^{-1}(\{z\})} \begin{cases} \mathbb{P}[\,Y \simeq y\,] & \text{, for } \Omega = \{y\} \\ \mathbb{P}[\,Y \in [\,l,\, r\,)\,] & \text{, for } \Omega = [\,l,\, r\,) \end{cases}$$

where, again, $\mathbb{P}[\,Y \simeq y\,] = f_Y(y)\ \partial y$ where f_Y is PDF of Y.

Censoring occurs commonly in a few different conditions. The first, called Type I censoring, is that a fixed number of experiments are set up to observe the time when some event occurs. Each of the experiments is stopped after a certain time, and where the event has not occurred, the events time for those experiments will be right censored. The second, called Type II censoring, is when the number of events to be observed is fixed. In this case, the duration of the experiment is a random variable, which censors the remaining subjects from right. Meanwhile, interval censoring arises naturally from situations where the events occur continuously, but are only inspected periodically, rendering the exact times of the events unknown, but bounded by the times of inspection.

Truncation and censoring can occur simultaneously. Moreover, in both conditions, the regions where the events are observed need no to be constant, but can be random variables. Typically, in such case it can be assumed that the truncating and censoring variables are observed and are independent of the data variables, in which case the above formulation applies.

The truncated and censored data can be used in an ML estimator in a straightforward manner, as long as care is taken with the values which are exact and which are censored. Let X_i be iid random variables, first truncated by the interval $B_i = [\, l_{B_i},\, r_{B_i}\,)$ and then censored into the interval $A_i = [\, l_{A_i},\, r_{A_i}\,)$. In this case, the experimenter observes the realizations of (A_i, B_i) rather than X_i. The average log-likelihood $\bar{\ell}$ is equal up to a constant to:

$$\tilde{\ell} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\log(\, F_X(r_{A_i}) - F_X(l_{A_i})\,)}_{T_{A_i}} - \underbrace{\log(\, F_X(r_{B_i}) - F_X(l_{B_i})\,)}_{T_{B_i}}$$

with $T_{A_i} = \log(\, f_X(l_{A_i})\, \partial x\,)$ if A_i is exact, and $T_{B_i} = 0$ if there is no truncation by B_i. In this case, the ordinary likelihood is recovered, as required. Here, F_X is the cumulative distribution function (CDF) of X. The discrepancy between the average log-likelihood $\bar{\ell}$ and the above quantity $\tilde{\ell}$ is due to the fact that (cumulative) probability and density are of different units. However, this does not cause problems for ML estimation, as the constant only depends on ∂x, which is constant with respect to the parameters $\boldsymbol{\theta}$.

Since the average log-likelihood involves terms, which contain a difference of the CDFs inside a logarithm, the direct solution of this problem might be intractable even if F_X is of a simple form. In such case, it is possible to solve the censored problem by using the EM algorithm (Dempster et al., 1977; Section 4.2) with the latent variables representing the true values of the censored data, as long as the uncensored problem remains directly solvable.

The following example demonstrates truncated and censored data sets, their properties, and parameter estimation using data from MC simulations. Here, a simulation consists of generating n failure times such that they are iid exponentially distributed with a mean of 1. The measurement time is 1, and the if the failure does not occur in the time window $[0, 1)$, the sample is either (right) truncated or censored (Type I censoring). For the purposes of demonstration, the right censored values are censored to unity where applicable.

Such data is exemplified in Figure 4.2. For the simulations 10^6 samples was generated. In the figure, the upper panel shows examples of the data: samples marked with red crosses indicate exactly observed data, which will be available in both truncated and censored data sets, while the blue circles indicate the right censored data, which is available only for the censored data set. The middle panel shows the resulting histograms, and the PDF of the model. This demonstrates that with truncation, the probability mass of the unobserved events is redistributed to the observable region, while with censoring, the mass is placed at the censoring value (in terms of probability density, this value is infinite). Lower panel shows the CDF of the exact observations, a Kaplan-Meier estimate of the CDF using the censored data, and the true CDF. Note that while the Kaplan-Meier estimator is a good estimate of the model CDF, it cannot estimate the CDF outside of the observed range.

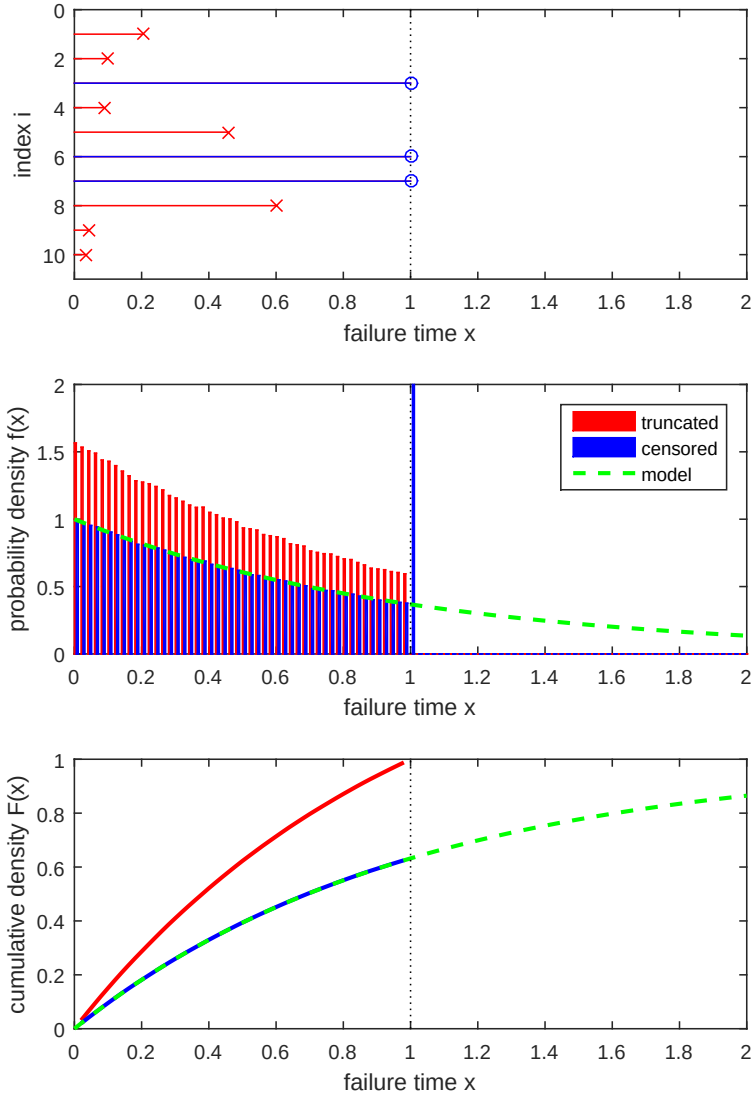

Figure 4.2: Upper panel: 10 samples of exponentially distributed data from MC simulations: 7 of the samples were exactly observed (red), while the samples 3, 6, and 7 remain right censored by the limited measurement time (blue). Middle panel: histograms of truncated and censored data (red and blue, respectively), and the PDF of the true model (green). Lower panel: CDF of the truncated data (red), the Kaplan-Meier estimate of the censored data (blue), and the CDF of the true model (green). The black vertical line represents the censoring time.

In this setting, on average, about e^{-1} (about 0.3679) of the samples will not be exactly observed. Consequently, in the case of truncation the data set contains fewer samples, while in the case of censoring, this quantity represents the fraction of right censored samples. The expected mean of the truncated data is $1-(e-1)^{-1}$ (about 0.4180), while the expected mean of the data in the censored data set is $1 - e^{-1}$ (about 0.6321), which is expected, since the large samples are either missing or underestimated.

Clearly, computing the sample means of the data sets to estimate the scale (i.e. the mean parameter of the model) of the model results in underestimation of the average failure time, in both cases. As the parameter is not known, it is not possible to correct this bias by e.g. subtracting it. However, it is possible to (asymptotically) correctly estimate the mean in both cases, by using the appropriate distributions. If all the values could be observed, the ML estimator for the non-censored data would be the sample mean:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where x_i are the n samples. Meanwhile, it can be shown that the ML estimator for the truncated data is the root of:

$$\mu - \frac{1}{\exp(1/\mu) - 1} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

where $(x_1, \cdots, x_m)$ are the exactly observed samples. The root exists, and is unique iff the sample mean is no larger than $1/2$. Otherwise, the ML estimator does not exist. Meanwhile, the estimator for the censored data is:

$$\hat{\mu} = \frac{1}{m} \left(\sum_{i=1}^{m} x_i + (n - m) \right)$$

where the term $(n - m)$ represents the sum of censoring times for the censored samples $(x_{m+1}, \cdots, x_n)$. Note that in the truncated and censored cases, the ML estimators do not exist for finite samples, if not enough exact observations are observed. This tends to be quite common: imagine observing a single right censored value – with the lack of any further evidence, what should be the estimate for the average? The above suggests $\hat{\mu} \to \infty$.

Next, it is of interest to analyze the performance of such estimators. If all the samples were exactly observed, the Fisher information of a observed sample about the mean parameter would be 1. Meanwhile, the Fisher information of a truncated sample about the mean is $1 - e\,(e - 1)^{-2}$ (about 0.0793), and the information of a potentially censored sample is $1 - e^{-1}$ (about 0.6321). In addition, for a fixed number of events, the truncated case has fewer samples available, so the estimator variance must be further scaled by $(1 - e^{-1})^{-1}$. Consequently, the

expected asymptotic variances of the ML estimators are about $1/n$, $19.9426/n$, and $1.5820/n$ for the uncensored, truncated, and censored data, respectively. Note that n represents the number of events, rather than samples, as above. An example of distribution of estimates demonstrating these differences in the estimator variance is shown in Figure 4.3. The distributions were generated using 10^6 simulations with 10^3 samples each. The results indicate that each method produces estimates around the true parameter value, but there are larger variations in the produced estimates when the data is truncated or censored. As expected, the results are more accurate when censoring is used instead of truncation.

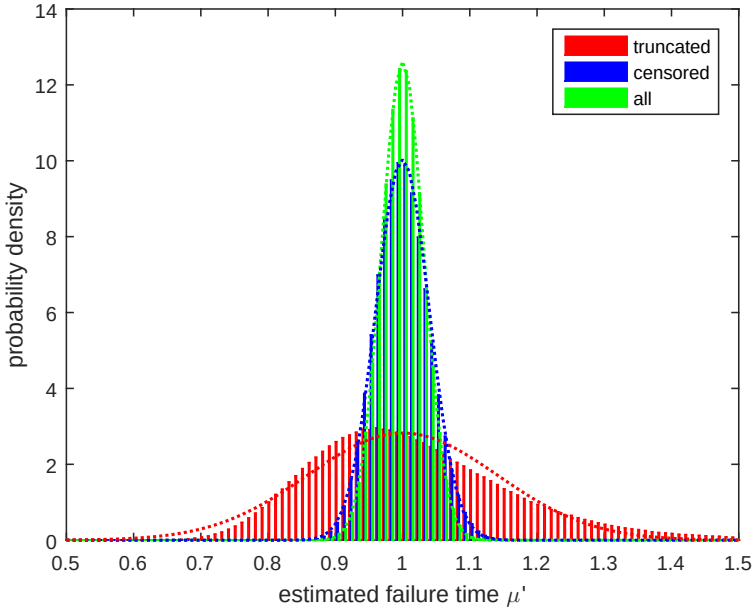

Figure 4.3: Histograms of estimates (bars) using truncated, censored, or all data (for the purposes of comparison). Each estimate was obtained by simulating 10^3 experiments. The dashed lines represent the expected distributions, which result from the asymptotic properties of the ML estimator.

4.4.2 Nonparametric estimation

Since the distribution of truncated and censored data differ from the underlying distribution, as demonstrated in the previous section, visualizing the data using histograms or empirical CDFs does not provide an accurate view of the underlying distribution. For this purpose, nonparametric estimators for the features of the underlying distribution have been developed.

One such estimator is the Kaplan-Meier estimator (Kaplan and Meier, 1958). This estimator estimates the survival function (or alternatively the CDF, as the two are intimately related) and can accommodate for uncensored and right censored

data (and possibly left truncation). Technically, it is the ML estimator of the survival function with the assumption that the function is piecewise constant. As expected, when there is no censored data, it degenerates to the empirical CDF.

The Kaplan-Meier estimator for the survival function is:

$$\hat{S}(x) = \prod_{x_i \leq x} \left(1 - \frac{d_i}{n_i}\right)$$

where x_i are the sorted observations $x_1 < \cdots < x_n$, n_i are the number of observations "at risk", that is, the observations which have not occurred or have been censored prior to time x_i, and d_i are the "deaths", the number of events occurring exactly at time x_i. Alternatively, the condition $x_i < x$ can be used in the product, making the survival function (and the CDF) left-continuous, as opposed to the right-continuous nature of the above definition.

A limitation of the Kaplan-Meier estimator is that it can only accommodate exact observations and right censoring. Turnbull's estimator is an extension of the Kaplan-Meier estimator, which can accommodate arbitrary modes of truncation and censoring (Turnbull, 1976). The Kaplan-Meier estimator is recovered in the case where only exact and right censored data is present. Like the Kaplan-Meier estimator, Turnbull's estimator is an ML estimator of the decrements of the survival function at the boundaries of the regions where the samples occur.

Turnbull's estimator cannot be expressed in closed form, but it can be found using the following EM algorithm. Consider n samples, where the sample x_i is truncated to the set B_i and then censored into the set A_i. First, the sample space is partitioned into m disjoint intervals $[q_i, p_i)$, such that each observation A_i can be formed as the union of such intervals. As the behavior of the survival function in the intervals $[q_j, p_j)$ does not affect the likelihood, the estimate can be specified using the increments $s_j = S(q_j) - S(p_j)$ in each interval. The latent variables are $I_{i,j}$, with $I_{i,j} = 1$ iff the sample x_i is in the interval $[q_j, p_j)$ and zero otherwise, and $J_{i,j}$ which represents the number of truncated (unobserved) samples in the interval $[q_j, p_j)$ for the observed sample x_i. This results in the following EM iteration: the E-step is:

$$\mu_{i,j}(\mathbf{s}) \doteq \mathbb{E}[\, I_{i,j} \mid A_i, B_i, s_j \,] = \delta\{\, [\, q_j, \, p_j \,) \subseteq A_i \,\}\, s_j \,/ \sum_{k=1}^{m} \delta\{\, [\, q_k, \, p_k \,) \subseteq A_i \,\}\, s_k$$

$$\nu_{i,j}(\mathbf{s}) \doteq \mathbb{E}[\, J_{i,j} \mid A_i, B_i, s_j \,] = \delta\{\, [\, q_j, \, p_j \,) \not\subseteq B_i \,\}\, s_j \,/ \sum_{k=1}^{m} \delta\{\, [\, q_k, \, p_k \,) \subseteq B_i \,\}\, s_k$$

and the M-step:

$$s_j^{(p+1)} = \sum_{i=1}^{n} \left(\mu_{i,j}(\mathbf{s}^{(p)}) + \nu_{i,j}(\mathbf{s}^{(p)})\right) \,/\, \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\mu_{i,j}(\mathbf{s}^{(p)}) + \nu_{i,j}(\mathbf{s}^{(p)})\right)$$

where $\delta\{\cdot\}$ is the indicator function.

Finally, the estimate for the survival function can be constructed as:

$$\hat{S}(x) = 1 - \sum_{q_j \leq x} s_j$$

where the behavior in each interval $[q_j, p_j)$ can be freely chosen. The intervals of ambiguity correspond to the ambiguity of the empirical CDF at the observations or the ambiguity of the Kaplan-Meier estimator beyond the last point if it is censored.

5 Conclusions and discussion

The objective of this thesis was to establish methods, which allow better characterization of the dynamics of bacterial genes and genetic networks from live cell measurement data. For this, novel computational methods were developed, which allow a more accurate and fine grained statistical quantification of RNA numbers and transcriptional dynamics from single-cell, single-molecule measurements. In each step, the performance of the methods was evaluated using both simulations and live cell measurement data. Finally, simulations were used to explore the nature and degree of the consequences of such dynamical differences, as quantifiable using the developed methods, on genetic networks.

Specifically, the methods developed in **Publications I** and **II** allow a standard quantification of RNA (and protein) numbers in live cells. The methods developed in the former publication were designed for static images, while that of the latter exploit temporal correlations for greater accuracy and to provide robustness against missing objects. The methods developed in **Publication III** are statistical methods for estimating the dynamical parameters of the subprocesses of bacterial transcription, designed for genome-wide quantification under a wide range of conditions. Such a statistical method is required, as the subprocesses cannot be directly measured in live cells. Finally, in **Publication IV**, the effects of changes in transcriptional dynamics, both of the rate and of the shape (stochasticity), on small genetic motifs were explored. In this work, two common motifs were considered: one of performing filtering in the amplitude and the other in the frequency domain. These results contribute in understanding how the behavior of small cellular circuits is affected by their constituent genes, particular by their transcriptional dynamics.

The methods developed in **Publication I** allow quantification of RNA numbers from the fluorescence intensities extracted from either RNA spots or from the total intensities inside each cell. The methods do not impose any distribution of the molecules to be quantified, but are based on a model of how the intensities are generated, which can be derived using the central limit theorem (CLT). The methods feature a numerical maximum likelihood (ML) estimator (based on the expectation maximization (EM) algorithm) of the intensity distribution, followed by a maximum posterior (MAP) classifier. For the estimator, two

variants are presented, one of which is expected to perform better, and the other which is computationally less expensive. The methods are fully automatic, as each parameter is estimated from the data. This allows objective comparison between independent measurements, as there are no tunable threshold parameters that could vary between the conditions, and facilities the large-scale analysis of measurement data.

The performance of the methods was analyzed using Monte Carlo (MC) simulations. The simulations suggest that biases are negligible and that the method is expected to perform well (above 80% accuracy) with signal-to-noise ratio (SNR), RNA distribution, and number of samples typical for single-RNA dynamics studies in *Escherichia coli*. It was also shown that the methods are more accurate than the previous method used for quantification of the RNA numbers in MS2-GFP-tagged RNA measurements (Golding et al., 2005). The new method performs significantly better when the RNA distribution is non-uniform, or the measurement noise is high (low SNR), and when neither is true, the method converges to the previous method. In addition, the methods allow estimating the expected accuracy of the quantification.

Moreover, the applicability of the methods was demonstrated by analyzing both spot and cell intensities extracted from live *E. coli* measurements. Analysis of spot intensities is expected to result in more accurate quantification, but require an accurate spot detection method; however, it was demonstrated the new method can work reliably in either mode. Finally, constructs with varying number of GFP tags were used (48 versus 96 binding sites), which demonstrates that the method can automatically adapt to various conditions, as designed.

Theoretically, the methods can be applied to any intensity distributions of fluorescent-tagged molecules. For example, low-expression level fluorescent proteins (e.g. tsr-Venus of Yu et al., 2006) are potential candidates in the future. Initial tests with such probes indicated that the SNR was too low for consistent accurate quantification, except for particular measurement conditions, requiring further optimization; however, the methods are likely to become more useful as the techniques of fluorescent tagging and microscopy improve.

Meanwhile, in **Publication II**, methods were developed to allow quantification of both RNA numbers and RNA production intervals from spot (or cell) intensity time series. Compared to the RNA quantification methods developed in **Publication I**, the new methods exploit the temporal information (that is, correlations between the RNA numbers) in the time series data to offer improved accuracy of the quantification. In addition, no pre- or post-processing is needed to extract the inter-transcription time intervals, which benefits studies of transcriptional dynamics (as opposed to RNA dynamics).

The developed methods consist of three steps: first, the data is grouped, distilling the temporal information; next, the parameters of the intensity model are estimated; and finally, a joint MAP classifier is used to find the globally optimal

estimate for the RNA numbers of the whole time series. An important feature of the method is that no regularization is applied in the temporal domain, such that time intervals are not biased. Also, as for the previous method, no model for the RNA distributions is implied. Two variants of the method were developed: one using a model similar to the previous method, derived using the CLT, and another, being a robust version of the former, designed to counter e.g. the RNA spots transiently leaving the focal plane.

Again, MC simulations were used to demonstrate the improved accuracy of the developed methods in comparison with two previous methods, both in estimating the RNA numbers and the production intervals. Further, MS2-GFP-tagged RNA measurements in live *E. coli* cells were used to demonstrate the applicability of the methods on measurement data. The methods are also applicable for purposes other than quantifying the MS2-GFP-tagged RNAs, provided that molecules are present in low copy numbers, their intensities come in distinct quanta, and that they degrade slowly (the last assumption is easily relaxed, but in such case the method can no longer be used for extracting the production intervals).

In **Publication III**, first a model was developed in order to combine the active-inactive promoter model (Peccoud and Ycart, 1995) with a sequential process of transcription initiation (McClure, 1985). Such a model is necessary for the analysis of transcription dynamics when performed in a genome-wide scale or in varying environmental conditions (Golding et al., 2005; Kandhavelu et al., 2012a; Muthukrishnan et al., 2012; Taniguchi et al., 2010; Yu et al., 2006). In addition to the model, methods for estimating its parameters in ML sense from inter-transcription interval data (e.g. extracted using the method of **Publication II**) were developed. The methods were designed to account for the finite and discrete nature of the measurement data. Finally, the methods also allow using statistical methods to identify the model components most responsible for the measured dynamical features.

The model and the methods were applied on both MC simulations and on measurement data of MS2-GFP-tagged RNA measurement data from live *E. coli*. Finally, it was demonstrated that the methods are applicable for other similar purposes by analyzing the kinetics of promoter activation. The methods enable genome-wide and differential analysis of the subprocesses of transcription initiation under a wide range of conditions. Further, it is expected that, with minor modifications, the methods can be extended to studies of eukaryotic transcription dynamics and of translational dynamics at the single protein level.

Finally, in **Publication IV**, MC simulations were used to investigate how the functioning of genetic filter motifs is affected by the changes in the dynamics of transcription initiation process of the constituent genes. For this, stochastic models of an amplitude filter and of a frequency filter were constructed, and SSA simulations were performed with various parameter sets in the ranges extracted from live *E. coli* measurements.

The genetic motifs constructed with stochastic gene expression models were found to differ significantly from their deterministic counterparts, suggesting that stochasticity plays a large role in the behavior of these circuits. Both the rate and the stochasticity of transcription initiation were found to affect the motifs in an intricate manner. Namely, the cutoffs of both filters are controlled by both features. Meanwhile, excess stochasticity was found to result in malfunctioning of the filters, particularly on the transition band. However, it was found that the adverse effects of low copy number fluctuations resulting from low expression levels can be mitigated by tuning the noise in the transcription initiation process. As such, one would expect that naturally occurring genes involved in filters with sensitive cutoffs have either high expression rates or employ a multistep transcription initiation process in order to constrain the noise in their RNA and protein numbers. Finally, the results demonstrated that the dynamical changes in the ranges identified by live cell measurements (Kandhavelu et al., 2012a; Muthukrishnan et al., 2012) are likely to have effects on the performance and design parameters of such motifs, suggesting that dynamical changes of transcription initiation, in the measured ranges, propagate to the level of gene networks.

Recently developed single-molecule measurement techniques in live cells (Pitchiaya et al., 2014) require carefully designed statistical methods for accurate and unbiased quantification and comparison of the results. General methods are unlikely candidates for such purpose: they may feature hidden assumptions which hinder the objective quantification of the studied properties, may fail to operate at a sufficient accuracy in a consistent manner, or may lack robustness against unexpected errors propagating from the earlier stages of the analysis. Accurate quantification is necessary, as even subtle changes may have implications on the level of genetic networks, as demonstrated in one of the publications.

The methods developed in this thesis feature greater accuracy of quantification than the previous methods, and allow quantifying features which were previously not possible. The improved quality of the results contributes in more objective characterization and comparison of the underlying phenomena, and enables studying the processes of gene expression in finer time scales. In addition, the developed methods allow a statistical analysis of the results, such as estimation of confidence or testing statistical hypotheses on the acquired results. Such advances are required in order to generate new insight on the dynamics and on the regulatory mechanisms of gene expression in *E. coli*.

As such, the methods will be critical for the success of single-cell, single-molecule studies which aid in understanding how changes in gene expression patterns are reflected on the behavior of the cells. Similar methods are currently being used to study the effects and mechanisms of activators and repressors (Kandhavelu et al., 2012a; Makela et al., 2013; Muthukrishnan et al., 2012) and environmental factors, such as stress conditions and temperature, (Muthukrishnan et al., 2012, 2014) on the expression of individual genes. Also, such methods have been used

to study how macromolecules are partitioned at cell division (Lloyd-Price et al., 2012), which has implications on cellular aging. These studies, now equipped with more capable tools, contribute in the understanding of cellular aging and diseases, and in the development of artificial genetic circuits, which can e.g. control the production of desired chemical compounds in the cells.

As demonstrated in some of the publications, the developed methods have potential for further applications on problems of similar nature, as they are adaptive by design. Also, the methods are likely applicable for other fluorescent tags and molecules, and novel applications are expected to arise when the techniques of GFP tagging and fluorescence microscopy are further developed. Finally, the methods can serve as a basis for constructing methods and tools for other similar problems, such as for quantifying eukaryotic transcription and the dynamics of translation.

Bibliography

Acar, M., Mettetal, J. T., and van Oudenaarden, A., "Stochastic switching as a survival strategy in fluctuating environments," *Nat. Genet.*, vol. 40, no. 4, pp. 471 – 475, 2008.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walters, P., *Molecular biology of the cell*, 6th ed. New York, NY: Garland Science, 2014.

Aldrich, J., "R. A. Fisher and the making of maximum likelihood," *Stat. Sci.*, vol. 13, no. 3, pp. 162 – 176, 1997.

Arkin, A., Ross, J., and McAdams, H. H., "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells," *Genetics*, vol. 149, no. 4, pp. 1633 – 1648, 1998.

Bai, L., Santangelo, T. J., and Wang, M. D., "Single-molecule analysis of RNA polymerase transcription," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 35, no. 1, pp. 343 – 360, 2006.

Becskei, A. and Serrano, L., "Engineering stability in gene networks by autoregulation," *Nature*, vol. 405, no. 6786, pp. 590 – 593, 2001.

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N., "Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 15, pp. 9697 – 9702, 2002.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., and Long, R. M., "Localization of *ASH1* mRNA particles in living yeast," *Mol. Cell.*, vol. 2, no. 4, pp. 437 – 445, 1998.

Blake, W. J., Kaern, M., Cantor, C. R., and Collins, J. J., "Noise in eukaryotic gene expression," *Nature*, vol. 422, no. 6932, pp. 633 – 637, 2003.

Bratsun, D., Volfson, D., Tsimring, L. S., and Hasty, J., "Delay-induced stochastic oscillations in gene regulation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 41, pp. 14 593 – 14 598, 2005.

Bremer, H., Dennis, P., and Ehrenberg, M., "Free RNA polymerase and modeling global transcription in *Escherichia coli*," *Biochimie*, vol. 85, no. 6, pp. 597–609, 2003.

Burgess, R. R., Travers, A. A., Dunn, J. J., and Bautz, E. K. F., "Factor stimulating transcription by RNA polymerase," *Nature*, vol. 221, no. 5175, pp. 43–46, 1969.

Cairns, B. R., "The logic of chromatin architecture and remodelling at promoters," *Nature*, vol. 461, no. 7261, pp. 193–198, 2009.

Campbell, R. E., Tour, O., Palmer, A. E., Steinbach, P. A., Baird, G. S., Zacharias, D. A., and Tsien, R. Y., "A monomeric red fluorescent protein," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 12, pp. 7877–7882, 2002.

Choi, P. J., Cai, L., Frieda, K., and Xie, X. S., "A stochastic single-molecule event triggers phenotype switching," *Science*, vol. 322, no. 5900, pp. 442–446, 2008.

Chong, S., Chen, C., Ge, H., and Xie, X. S., "Mechanism of transcriptional bursting in bacteria," *Cell*, vol. 158, no. 2, pp. 314–326, 2014.

Chowdhury, S., Kandhavelu, M., Yli-Harja, O., and Ribeiro, A. S., "Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE)," *BMC Bioinf.*, vol. 14, no. Suppl. 10, p. S8, 2013.

Chubb, J. R. and Liverpool, T. B., "Bursts and pulses: Insights from single cell studies into transcriptional mechanisms," *Curr. Opin. Genet. Dev.*, vol. 20, no. 5, pp. 478–484, 2006.

Chubb, J. R., Trcek, T., Shenoy, S. M., and Singer, R. H., "Transcriptional pulsing of a developmental gene," *Curr. Biol.*, vol. 16, no. 10, pp. 1018–1025, 2006.

Coelho, M., Maghelli, N., and Tolic-Norrelykke, I. M., "Single-molecule imaging in vivo: The dancing building blocks of the cell," *Integr. Biol.*, vol. 5, no. 5, pp. 748–758, 2013.

Cormack, B. P., Valdivia, R. H., and Falkow, S., "FACS-optimized mutants of the green fluorescent protein (GFP)," *Gene*, vol. 173, no. 1, pp. 33–38, 1996.

Crick, F., "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

Croucher, N. J. and Thomson, N. R., "Studying bacterial transcriptomes using RNA-seq," *Curr. Opin. Microbiol.*, vol. 13, no. 5, pp. 619–624, 2010.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K., "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE T. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B Met.*, vol. 39, no. 1, pp. 1–38, 1977.

Elowitz, M. B. and Leibler, S., "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.

Elowitz, M. B., Surette, M. G., Wolf, P.-E., Stock, J. B., and Leibler, S., "Protein mobility in the cytoplasm of *Escherichia coli*," *J. Bacteriol.*, vol. 181, no. 1, pp. 197–203, 1999.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S., "Stochastic gene expression in a single cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J., and Eisen, M. B., "Noise minimization in eukaryotic gene expression," *PLoS Biol.*, vol. 2, no. 6, p. e137, 2004.

Fusco, D., Accornero, N., Lavoie, B., Shenoy, S. M., Blanchard, J.-M., Singer, R. H., and Bertrand, E., "Single mRNA molecules demonstrate probabilistic movement in living mammalian cells," *Curr. Biol.*, vol. 13, no. 2, pp. 161–167, 2003.

Gardner, T. S., Cantor, C. R., and Collins, J. J., "Construction of a genetic toggle switch in *Escherichia coli*," *Nature*, vol. 403, no. 6767, pp. 339–342, 2000.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S., "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–741, 2003.

Gillespie, D. T., "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J. Comput. Phys.*, vol. 22, no. 4, pp. 403–434, 1976.

Gillespie, D. T., "Concerning the validity of the stochastic approach to chemical kinetics," *J. Stat. Phys.*, vol. 16, no. 3, pp. 311–318, 1977.

Gillespie, D. T., "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, no. 25, pp. 2340–2361, 1977.

Gillespie, D. T., "A rigirious derivation of chemical master equation," *Physica A*, vol. 188, no. 1–3, pp. 404–425, 1992.

Gillespie, D. T., "The chemical Langevin equation," *J. Chem. Phys.*, vol. 113, no. 1, pp. 297–306, 2000.

Gillespie, D. T., "Approximate accelerated stochastic simulation of chemically reacting systems," *J. Chem. Phys.*, vol. 115, no. 1, pp. 1716–1733, 2001.

Gillespie, D. T., "Stochastic simulation of chemical kinetics," *Annu. Rev. Phys. Chem.*, vol. 58, no. 1, pp. 35–55, 2007.

Gillespie, D. T., "Deterministic limit of stochastic chemical kinetics," *J. Phys. Chem. B*, vol. 113, no. 6, pp. 1640–1644, 2009.

Golding, I. and Cox, E. C., "RNA dynamics in live *Escherichia coli* cells," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 31, pp. 11 310–11 315, 2004.

Golding, I. and Cox, E. C., "Eukaryotic transcription: What does it mean for a gene to be 'on'?" *Curr. Biol.*, vol. 16, no. 10, pp. R371–R373, 2006.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C., "Real-time kinetics of gene activity in individual bacteria," *Cell*, vol. 123, no. 6, pp. 1025–1036, 2005.

Goldman, S. R., Ebright, R. H., and Nickels, B. E., "Direct detection of abortive RNA transcripts in vivo," *Science*, vol. 324, no. 5929, pp. 927–928, 2009.

Greive, S. J. and von Hippel, P. H., "Thinking quantitatively about transcriptional regulation," *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 3, pp. 221–232, 2005.

Gries, T. J., Kontur, W. S., Capp, M. W., Saecker, R. M., and Record, Jr., M. T., "One-step DNA melting in the RNA polymerase cleft opens the initiation bubble to form an unstable open complex," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, no. 23, pp. 10 418–10 423, 2010.

Gronlund, A., Lotstedt, P., and Elf, J., "Transcription factor binding kinetics constrain noise suppression via negative feedback," *Nat. Commun.*, vol. 4, p. 1864, 2013.

Grundy, F. J. and Henkin, T. M., "From ribosome to riboswitch: Control of gene expression in bacteria by RNA structural rearrangements," *Crit. Rev. Biochem. Mol. Biol.*, vol. 41, no. 6, pp. 329–338, 2006.

Guptasarma, P., "Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of *Escherichia coli*?" *Bioessays*, vol. 17, no. 11, pp. 987–997, 1995.

Hakkinen, A. and Ribeiro, A. S., "Characterizing rate limiting steps in transcription from RNA production times in live cells," *Bioinformatics*, in press, doi: 10.1093/bioinformatics/btv744, 2015.

Hakkinen, A. and Ribeiro, A. S., "Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data," *Bioinformatics*, vol. 31, no. 1, pp. 69–75, 2015.

Hakkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M., and Ribeiro, A. S., "CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*," *Bioinformatics*, vol. 29, no. 3, pp. 1708 – 1709, 2013.

Hakkinen, A., Tran, H., Yli-Harja, O., and Ribeiro, A. S., "Effects of rate-limiting steps in transcription initiation on genetic filter motifs," *PLoS One*, vol. 8, no. 8, p. e70439, 2013.

Hakkinen, A., Kandhavelu, M., Garasto, S., and Ribeiro, A. S., "Estimation of fluorescence-tagged RNA numbers from spot intensities," *Bioinformatics*, vol. 30, no. 8, pp. 1146 – 1153, 2014.

Hastings, W. K., "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97 – 109, 1970.

Herbert, K. M., La Porta, A., Wong, B. J., Mooney, R. A., Neuman, K. C., Landick, R., and Block, S. M., "Sequence-resolved detection of pausing by single RNA polymerase molecules," *Cell*, vol. 125, no. 6, pp. 1083 – 1094, 2006.

Hill, A. V., "The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves," *J. Physiol.*, vol. 40, no. Suppl., pp. iv – vii, 1910.

Hotelling, H., "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417 – 441, 1933.

Hsu, L. M., "Promoter clearance and escape in prokaryotes," *BBA Gene Struct. Expr.*, vol. 1577, no. 2, pp. 191 – 207, 2002.

Huh, D. and Paulsson, J., "Non-genetic heterogeneity from stochastic partitioning at cell division," *Nat. Genet.*, vol. 43, no. 2, pp. 95 – 100, 2011.

Iglesias, P. A. and Ingalls, B. P., *Control theory and systems biology.* Cambridge, MA: MIT Press, 2009.

Johansson, H. E., Dertinger, D., LeCuyer, K. A., Behlen, L. S., Greef, C. H., and Uhlenbeck, O. C., "A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, no. 16, pp. 9244 – 9249, 1998.

Kabata, H., Kurosawa, O., Arai, I., Washizu, M., Margarson, S. A., Glass, R. E., and Shimamoto, N., "Visualization of single molecules of RNA polymerase sliding along DNA," *Science*, vol. 262, no. 5139, pp. 1561 – 1563, 1993.

Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J., "Stochasticity in gene expression: From theories to phenotypes," *Nat. Rev. Genet.*, vol. 6, no. 6, pp. 451 – 464, 2005.

Kandhavelu, M., Mannerstrom, H., Gupta, A., Hakkinen, A., Lloyd-Price, J., Yli-Harja, O., and Ribeiro, A. S., "In vivo kinetics of transcription initiation of the lar promoter in *Escherichia coli*: Evidence for a sequential mechanism with two rate-limiting steps," *BMC Syst. Biol.*, vol. 5, p. 149, 2011.

Kandhavelu, M., Hakkinen, A., Yli-Harja, O., and Ribero, A. S., "Single-molecule dynamics of transcription of the lar promoter," *Phys. Biol.*, vol. 9, no. 2, p. 026004, 2012.

Kandhavelu, M., Lloyd-Price, J., Gupta, A., Muthukrishnan, A.-B., Yli-Harja, O., and Ribeiro, A. S., "Regulation of mean and noise of the in vivo kinetics of transcription under the control of the *lac/ara-1* promoter," *FEBS Lett.*, vol. 586, no. 21, pp. 3870 – 3875, 2012.

Kaplan, E. L. and Meier, P., "Nonparametric estimation from incomplete observations," *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457 – 481, 1958.

Klumpp, S. and Hwa, T., "Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 47, pp. 18 159 – 18 164, 2008.

Kurtz, T. G., "The relationship between stochastic and deterministic models for chemical reactions," *J. Chem. Phys.*, vol. 57, no. 7, pp. 2976 – 2978, 1972.

Landick, R., "The regulatory roles and mechanism of transcriptional pausing," *Biochem. Soc. Trans.*, vol. 34, no. 6, pp. 1062 – 1066, 2006.

Lawless, J. F., *Statistical models and methods for lifetime data*, 2nd ed. Hoboken, NJ: Wiley, 2003.

Lehmann, E. L. and Casella, G., *Theory of point estimation*, 2nd ed. New York, NY: Springer-Verlag, 1998.

Lestas, I., Vinnicombe, G., and Paulsson, J., "Fundamental limits on the suppression of molecular fluctuations," *Nature*, vol. 467, no. 7312, pp. 174 – 178, 2010.

Lindner, A. B., Madden, R., Demarez, A., Stewart, E. J., and Taddei, F., "Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 8, pp. 3076 – 3081, 2008.

Lloyd, S. P., "Least squares quantization in PCM," *IEEE T. Inform. Theory*, vol. 28, no. 2, pp. 129 – 137, 1957.

Lloyd-Price, J., Lehtivaara, M., Kandhavelu, M., Chowdhury, S., Muthukrishnan, A.-B., Yli-Harja, O., and Ribeiro, A. S., "Probabilistic RNA partitioning generates transient increases in the normalized variance of RNA numbers in

synchronized populations of *Escherichia coli*," *Mol. BioSyst.*, vol. 8, no. 2, pp. 565 – 571, 2012.

Locke, J. C. W. and Elowitz, M. B., "Using movies to analyse gene circuit dynamics in single cells," *Nat. Rev. Microbiol.*, vol. 7, no. 5, pp. 383 – 392, 2009.

Lutz, R. and Bujard, H., "Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements," *Nucl. Acids Res.*, vol. 25, no. 6, pp. 1203 – 1210, 1997.

Lutz, R., Lozinski, T., Ellinger, T., and Bujard, H., "Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator," *Nucl. Acids Res.*, vol. 29, no. 18, pp. 3873 – 3881, 2001.

Makela, J., Lloyd-Price, J., Yli-Harja, O., and Ribeiro, A. S., "Stochastic sequence-level model of coupled transcription and translation in prokaryotes," *BMC. Bioinf.*, vol. 122, no. 1, pp. 1471 – 2105, 2011.

Makela, J., Kandhavelu, M., Oliveira, S. M. D., Chandraseelan, J. G., Lloyd-Price, J., Peltonen, J., Yli-Harja, O., and Ribeiro, A. S., "*In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter," *Nucl. Acids Res.*, vol. 41, no. 13, pp. 6544 – 6552, 2013.

Mann, H. B. and Wald, A., "On stochastic limit and order relationships," *Ann. Math. Statist.*, vol. 14, no. 3, pp. 217 – 226, 1943.

Marsaglia, G., "Random numbers fall mainly in the planes," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 61, no. 1, pp. 25 – 28, 1968.

Martinez-Antonio, A. and Collado-Vides, J., "Identifying global regulators in transcriptional regulatory networks in bacteria," *Curr. Opin. Microbiol.*, vol. 6, no. 5, pp. 482 – 489, 2006.

Matsumoto, M. and Nishimura, T., "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM T. Model. Comput. S.*, vol. 8, no. 1, pp. 3 – 30, 1998.

McAdams, H. H. and Arkin, A., "Stochastic mechanisms in gene expression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 94, no. 3, pp. 814 – 819, 1997.

McClure, W. R., "Rate-limiting steps in RNA chain initiation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 77, no. 10, pp. 5634 – 5638, 1980.

McClure, W. R., "Mechanism and control of transcription initiation in prokaryotes," *Annu. Rev. Biochem.*, vol. 54, pp. 171 – 204, 1985.

McQuarrie, D. A., "Stochastic approach to chemical kinetics," *J. Appl. Prob.*, vol. 4, no. 3, pp. 413 – 478, 1967.

Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., Kudasov, F., Chowdhury, S., Yli-Harja, O., and Ribeiro, A. S., "Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells," *Nucl. Acids Res.*, vol. 40, no. 17, pp. 8472–8483, 2012.

Muthukrishnan, A.-B., Martikainen, A., Neeli-Venkata, R., and Ribeiro, A. S., "*In vivo* transcription kinetics of a synthetic gene uninvolved in stress-response pathways in stressed *Escherichia coli* cells," *PLoS ONE*, vol. 9, no. 9, p. e109005, 2014.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S., "Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise," *Nature*, vol. 441, no. 7095, pp. 840–846, 2006.

Neyman, J. and Pearson, E. S., "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. R. Roy. Soc. Lond. A*, vol. 231, no. 694–706, pp. 289–337, 1933.

Nudler, E. and Gottesman, M. E., "Transcription termination and anti-termination in *E. coli*," *Genes Cells*, vol. 7, no. 8, pp. 755–768, 2002.

Otsu, N., "A threshold selection method from gray-level histograms," *IEEE Trans. Sys., Man., Cyber.*, vol. 9, no. 1, pp. 62–66, 1979.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A., "Regulation of noise in the expression of a single gene," *Nat. Genet.*, vol. 31, no. 1, pp. 69–73, 2002.

Park, S. K. and Miller, K. W., "Random number generators: Good ones are hard to find," *Commun. ACM*, vol. 31, no. 10, pp. 1192–1201, 1988.

Parzen, E., "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.

Paulsson, J., "Summing up the noise in gene networks," *Nature*, vol. 472, no. 6973, pp. 415–418, 2004.

Paulsson, J., Berg, O. G., and Ehrenberg, M., "Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, no. 13, pp. 7148–7153, 2000.

Peabody, D. S. and Lim, F., "Complementation of RNA binding site mutations in MS2 coat protein heterodimers," *Nucl. Acids Res.*, vol. 24, no. 12, pp. 2352–2359, 1996.

Pearson, K., "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, no. 6, pp. 559–572, 1901.

Peccoud, J. and Ycart, B., "Markovian modeling of gene-product synthesis," *Theor. Popul. Biol.*, vol. 48, no. 2, pp. 222–234, 1995.

Pitchiaya, S., Heinicke, L. A., Custer, T. C., and Walter, N. G., "Single molecule fluorescence approaches shed light on intracellular RNAs," *Chem. Rev.*, vol. 114, no. 6, pp. 3224–3265, 2014.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S., "Stochastic mRNA synthesis in mammalian cells," *PLoS Biol.*, vol. 4, no. 10, p. e309, 2006.

Rajala, T., Hakkinen, A., Yli-Harja, O., and Ribeiro, A. S., "Effects of transcriptional pausing on gene expression dynamics," *PLoS Comput. Biol.*, vol. 6, no. 3, p. e1000704, 2010.

Raser, J. M. and O'Shea, E. K., "Control of stochasticity in eukaryotic gene expression," *Science*, vol. 304, no. 5678, pp. 1811–1814, 2004.

Record, Jr., M. T., Reznikoff, W. S., Craig, M. L., McQuade, K. L., and Schlax, P. J., "*Escherichia coli* RNA polymerase ($E\sigma^{70}$), promoters, and the kinetics of the steps of transcription initiation," in *Escherichia coli and Salmonella typhimurium: Cellular and molecular biology*, 2nd ed., Neidhart, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M., and Umbarger, H. E., Eds. Washington, DC: ASM Press, 1996, vol. 2.

Ribeiro, A. S., Zhu, R., and Kauffman, S. A., "A general modeling strategy for gene regulatory networks with stochastic dynamics," *J. Comput. Biol.*, vol. 13, no. 9, pp. 1630–1639, 2006.

Ribeiro, A. S., Smolander, O.-P., Rajala, T., Hakkinen, A., and Yli-Harja, O., "Delayed stochastic model of transcription at the single nucleotide level," *J. Comput. Biol.*, vol. 16, no. 4, pp. 539–553, 2009.

Richardson, J. P., "Rho-dependent termination and ATPases in transcript termination," *BBA Gene Struct. Expr.*, vol. 1577, no. 2, pp. 251–260, 2002.

Rosenblatt, M., "Remarks on some nonparametric estimates of a density function," *Ann. Math. Stat.*, vol. 27, no. 3, pp. 832–837, 1956.

Roussel, M. R. and Zhu, R., "Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression," *Phys. Biol.*, vol. 8, no. 3, pp. 274–284, 2006.

Ruusuvuori, P., Aijo, T., Chowdhury, S., Garmendia-Torres, C., Selinummi, J., Birbaumer, M., Dudley, A. M., Pelkmans, L., and Yli-Harja, O., "Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images," *BMC Bioinf.*, vol. 11, no. 1, p. 248, 2010.

Saecker, R. M., Record, Jr., M. T., and deHaseth, P. L., "Mechanism of bacterial transcription initiation," *J. Mol. Biol.*, vol. 412, no. 5, pp. 754 – 771, 2011.

Samoilov, M., Arkin, A., and Ross, J., "Signal processing by simple chemical systems," *J. Phys. Chem. A*, vol. 106, no. 43, pp. 10 205 – 10 221, 2002.

Samoilov, M., Plyasunov, S., and Arkin, A. P., "Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 7, pp. 2310 – 2315, 2005.

Schlax, P. J., Capp, M. W., and Record, M. Thomas, J., "Inhibition of transcription initiation by lac repressor," *J. Mol. Biol.*, vol. 245, no. 4, pp. 331 – 350, 1995.

Shahrezaei, V. and Swain, P. S., "Analytical distributions for stochastic gene expression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 45, pp. 17 256 – 17 261, 2008.

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U., "Network motifs in the transcriptional regulation network of *Escherichia coli*," *Nat. Genet.*, vol. 31, no. 1, pp. 64 – 68, 2002.

So, L., Ghosh, A., Zong, C., Sepulveda, L. A., Segev, R., and Golding, I., "General properties of transcriptional time series in *Escherichia coli*," *Nat. Genet.*, vol. 43, no. 6, pp. 554 – 560, 2011.

Storz, G. and Gottesman, S., "Versatile roles of small RNA regulators in bacteria," in *The RNA world*, 3rd ed., Gesteland, R. F., Cech, T. R., and Atkins, J. F., Eds. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2006, vol. 43.

Suel, G. M., Garcia-Ojalvo, J., Liberman, L. M., and Elowitz, M. B., "An excitable gene regulatory circuit induces transient cellular differentiation," *Nature*, vol. 440, no. 7083, pp. 545 – 550, 2006.

Suel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J., and Elowitz, M. B., "Tunability and noise dependence in differentiation dynamics," *Science*, vol. 315, no. 5819, pp. 1716 – 1719, 2007.

Sundberg, R., "Maximum likelihood theory for incomplete data from an exponential family," *Scand. J. Stat.*, vol. 1, no. 2, pp. 49 – 58, 1974.

Suzuki, T., Matsuzaki, T., Hagiwara, H., Aoki, T., and Takata, K., "Recent advances in fluorescent labeling techniques for fluorescence microscopy," *Acta Histochem. Cytochem.*, vol. 40, no. 5, pp. 131 – 137, 2007.

Swain, P. S., Elowitz, M. B., and Siggia, E. D., "Intrinsic and extrinsic contributions to stochasticity in gene expression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 20, pp. 12 795 – 12 800, 2002.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S., "Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells," *Science*, vol. 329, no. 5991, pp. 533 – 538, 2010.

Thattai, M. and van Oudenaarden, A., "Stochastic gene expression in fluctuating environments," *Genetics*, vol. 167, no. 1, pp. 523 – 530, 2004.

Turnbull, B. W., "The empirical distribution function with arbitrarily grouped, censored and truncated data," *J. Roy. Stat. Soc. B Met.*, vol. 38, no. 3, pp. 290 – 295, 1976.

Uptain, S. M., Kane, C. M., and Chamberlin, M. J., "Basic mechanisms of transcript elongation and its regulation," *Annu. Rev. Biochem.*, vol. 66, no. 1, pp. 117 – 172, 1997.

Van Trees, H. L., Bell, K. L., and Tian, Z., *Detection, estimation and modulation theory: Part I*, 2nd ed. Hoboken, NJ: Wiley, 2013.

Vogel, U. and Jensen, K. F., "The RNA chain elongation rate in *Escherichia coli* depends on the growth rate," *J. Bacteriol.*, vol. 176, no. 10, pp. 2807 – 2813, 1994.

von Hippel, P. H., "An integrated model of the transcription complex in elongation, termination, and editing," *Science*, vol. 281, no. 5377, pp. 660 – 665, 1998.

Walter, G., Zillig, W., Palm, P., and Fuchs, E., "Initiation of DNA-dependent RNA synthesis and the effect of heparin on RNA polymerase," *Eur. J. Biochem.*, vol. 3, no. 2, pp. 194 – 201, 1967.

Wang, F., Redding, S., Finkelstein, I. J., Gorman, J., Reichman, D. R., and Greene, E. C., "The promoter-search mechanism of *Escherichia coli* RNA polymerase is dominated by three-dimensional diffusion," *Nat. Struct Mol. Biol.*, vol. 20, no. 2, pp. 174 – 181, 2013.

Weinberg, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., and Schaffer, D. V., "Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity," *Cell*, vol. 122, no. 2, pp. 169 – 182, 2005.

Wilks, S. S., "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Stat.*, vol. 9, no. 1, pp. 60 – 62, 1938.

Wolf, D. M. and Arkin, A. P., "Motifs, modules and games in bacteria," *Curr. Opin. Microbiol.*, vol. 6, no. 2, pp. 125 – 134, 2003.

Wu, C. F. J., "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, no. 1, pp. 95 – 103, 1983.

Wu, J. Q. and Snyder, M., "RNA polymerase II stalling: Loading at the start prepares genes for a sprint," *Genome Biol.*, vol. 9, no. 5, p. 220, 2008.

Young, J. W., Locke, J. C. W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P. S., Mjolsness, E., and Elowitz, M. B., "Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy," *Nat. Protoc.*, vol. 7, no. 1, pp. 80 – 88, 2012.

Yu, J., Xiao, J., Run, X., Lao, K., and Xie, X. S., "Probing gene expression in live cells, one protein molecule at a time," *Science*, vol. 311, no. 5767, pp. 1600 – 1603, 2006.

Publications

Publication I

Hakkinen, A., Kandhavelu, M., Garasto, S., and Ribeiro, A. S., "Estimation of fluorescence-tagged RNA numbers from spot intensities," *Bioinformatics*, vol. 30, no. 8, pp. 1146 – 1153, 2014.

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1146–1153
doi:10.1093/bioinformatics/btt766

Gene expression

Advance Access publication January 2, 2014

Estimation of fluorescence-tagged RNA numbers from spot intensities

Antti Häkkinen, Meenakshisundaram Kandhavelu, Stefania Garasto and Andre S. Ribeiro*

Department of Signal Processing, Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group, Tampere University of Technology, Tampere, Finland

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Present research on gene expression using live cell imaging and fluorescent proteins or tagged RNA requires accurate automated methods of quantification of these molecules from the images. Here, we propose a novel automated method for classifying pixel intensities of fluorescent spots to RNA numbers.

Results: The method relies on a new model of intensity distributions of tagged RNAs, for which we estimated parameter values in maximum likelihood sense from measurement data, and constructed a maximum a posteriori classifier to estimate RNA numbers in fluorescent RNA spots. We applied the method to estimate the number of tagged RNAs in individual live *Escherichia coli* cells containing a gene coding for an RNA with MS2-GFP binding sites. We tested the method using two constructs, coding for either 96 or 48 binding sites, and obtained similar distributions of RNA numbers, showing that the method is adaptive. We further show that the results agree with a method that uses time series data and with quantitative polymerase chain reaction measurements. Lastly, using simulated data, we show that the method is accurate in realistic parameter ranges. This method should, in general, be applicable to live single-cell measurements of low-copy number fluorescence-tagged molecules.

Availability and implementation: MATLAB extensions written in C for parameter estimation and finding decision boundaries are available under Mozilla public license at http://www.cs.tut.fi/%7ehakkin22/estrna/.

Contact: andre.ribeiro@tut.fi

Received on October 28, 2013; revised on December 9, 2013; accepted on December 25, 2013

1 INTRODUCTION

The processes of production and degradation of RNA and protein numbers are stochastic, which causes these numbers to differ between individuals (Elowitz *et al*., 2002; Kaern *et al*., 2005; Ozbudak *et al*., 2002; Taniguchi *et al*., 2010; Yu *et al*., 2006) and to vary over time in individual cells, even under optimal stable conditions (Golding *et al*., 2005; Kandhavelu *et al*., 2012b). Present measurement techniques to study these processes in live cells rely on the usage of fluorescent proteins (Montero Llopis *et al*., 2010; Raj *et al*., 2008), and consequently on fluorescent microscopy and image processing methods, to extract the relevant statistical data.

One of the most accurate techniques to quantify gene expression dynamics *in vivo* consists of tagging target RNA molecules with multiple MS2d-GFP proteins, which makes the target RNA transcripts visible as bright spots (Golding and Cox, 2004), as soon as these are transcribed (Golding and Cox, 2004; Peabody, 1993). The present method of quantifying the RNA numbers in cells or in the fluorescent RNA spots relies on a manual selection of the intensity of a single-tagged RNA using the histogram of fluorescence intensities (Golding *et al*., 2005). Following this selection, this method assumes that the peaks of spots intensities are concentrated at multiples of the expected intensity of a single RNA (Golding *et al*., 2005), so as to quantify the number of RNA molecules in spots of varying fluorescence intensities.

This method of quantification is not optimal for several reasons. First, it relies on human intervention, and the manual selection can have a major effect in the results of the process of counting the RNA molecules. This hampers comparison between results because the selection of the intensity of the first peak varies between users. Second, the distribution of the number of RNA molecules composing the spots is not uniform, and thus the optimal quantification of these numbers is not achieved by simple rounding. Third, the variance in intensities of the spots is expected to increase with the number of tagged RNAs composing the spot. Fourth, the tagged RNA molecules tend to accumulate at the poles of the cells, and one cannot rely on spatial separation between them (Golding and Cox, 2004; Lloyd-Price *et al*., 2012). Lastly, for large datasets, manually assisted quantification can be excessively laborious.

Here, we propose an automatic method that improves on the RNA quantification from fluorescence images. For this, we establish a mathematical model of the intensities, which is free from any assumption on the shape of the distribution of the RNAs, as this distribution is the subject of the study. Also, we propose methods to estimate the parameters of the model and construct a classifier. We then exemplify the usage of the method in determining the number of RNAs in clusters of MS2-GFP-tagged RNA molecules in *Escherichia coli* under various conditions. In particular, we compare results of analyzing cells with one of two target RNA constructs that differ in the number of binding sites for MS2-GFP proteins. In addition, the results are compared with a method that uses temporal information for the RNA quantification, and results from cells subject to different

*To whom correspondence should be addressed.

levels of induction of target RNA are compared with quantitative polymerase chain reaction (qPCR) measurements. Lastly, we study the performance of the classifier on simulated data, whose ground truth is known, and compare it with that of the previous method.

2 SYSTEM AND METHODS

2.1 Cells and plasmids

Escherichia coli strain DH5α-PRO was provided by I. Golding (University of Illinois) and contains two constructs: a PROTET-K133 medium-copy vector carrying a MS2d-GFP reporter, controlled by $P_{LtetO-1}$, and the pIG-BAC single-copy vector coding for mRFP1-MS2-96bs RNA, whose expression is controlled by $P_{lac/ara-1}$ (Golding and Cox, 2004).

The second construct was engineered by us, and it was designed to contain 48 binding sites (bs) for MS2-GFP. The target-gene vector pMK-BAC, containing $P_{lac/ara-1}$ with a 48 MS2-GFP binding site array in a single-copy bacterial artificial chromosome (BAC), was constructed using standard molecular biology methods. $P_{lac/ara-1}$-48bs was amplified with smaI restriction endonuclease from a BAC clone carrying a target gene $P_{lac/ara-1}$-mRFP1-96bs. The primers (forward: 5′ CCCGGGGAAGACATGAGGATCA 3′ and reverse: 5′ CCCGG GTCAATTCTGTGTGAAATTG 3′) were designed to amplify the $P_{lac/ara-1}$-48bs with smaI restriction site flanking regions. The amplicon and the BAC vector were subjected to smaI restriction digestion, followed by ligation of the amplified product. We obtained a single-copy F-based plasmid carrying the target region $P_{lac/ara-1}$ with a 48bs array. This product was transformed into the competent *E.coli* strain DH5α-PRO. The recombinants were selected with antibiotic screening and confirmed with sequence analysis.

2.2 Microscopy measurements

Cells were grown in Miller lysogeny broth (LB) medium, supplemented with antibiotics according to the specific plasmids. Cells were grown overnight at 37°C with aeration, diluted into fresh medium and allowed to grow at 37°C until an optical density of OD600 of 0.3–0.5 was reached. To attain full induction of the MS2d-GFP reporter, cells were incubated with 100 ng/ml of anhydrotetracycline (aTc, from IBA GmbH). In all, 0.1% of L-arabinose (Sigma-Aldrich) and 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG, Fermentas) were used to fully induce the target RNA. In one case, IPTG was not added, so that the target gene remains only weakly induced. Cells were preincubated with arabinose at the same time as aTc. IPTG was added (if added) 1 h after aTc, and cells were incubated for 5 min.

Microscopy was performed using a Nikon Eclipse (TE-2000-U, Nikon, Tokyo, Japan) inverted confocal laser-scanning microscope. Cells were imaged in a thermal chamber set to 37°C. Single time-point images were taken 1 h after induction by IPTG (if induced). For time series, images were taken 5 min after induction by IPTG, for 2 h, once per minute.

For imaging, a few microliter of culture were placed between a coverslip and a slab of 1% agarose containing LB along with the appropriate concentrations of inducers. When both the reporter and the target RNA are present in the cells, MS2d-GFP proteins bind to the target RNA, forming a bright fluorescent spot (Golding *et al.*, 2005). The RNA becomes visible during or shortly after elongation (Golding and Cox, 2004).

2.3 The qPCR analysis of the target RNA

The target RNA was induced as described earlier in the text. Following induction, the cells were immediately fixed with RNAprotect bacteria reagent, followed by enzymatic lysis with Tris-EDTA lysozyme buffer (pH 8.3). From the lysed cells, total RNA was isolated with RNeasy RNA purification kit (Qiagen), according to the manufacturer's instructions. DNaseI treatment was performed to avoid DNA contamination. The complementary DNA (cDNA) was synthesized (Fermentas, Finland) from 1 μg of RNA with iScript Reverse Transcription Supermix, according to the manufacturer's instructions. The cDNA templates with final concentration of 10 ng/μl were added to the qPCR master mix, which contained iQ SYBR Green supermix (Fermentas, Finland) with primers for the target and reference genes at a final concentration of 200 nM.

We used the 16S ribosomal RNA housekeeping gene for internal reference. The primers for the target mRNA (forward: 5′ TACGACGCCGAGGTCAAG 3′ and reverse: 5′ TTGTGGGA GGTGATGTCCA 3′) target the mRFP1 coding region and the reference gene 16S ribosomal RNA (forward: 5′ CGTCAGCTCGTGTTGTGAA 3′ and reverse: 5′ GGACCGCTGGCAACAAAG 3′). The experiment was performed using a Biorad MiniOpticon real-time PCR system (Biorad, Finland) with the following thermal cycling protocol: 40 cycles of 95°C for 10 s, 52°C for 30 s and 72°C for 30 s for each cDNA replicate. Reactions were performed in two experiments, each with two replicates per condition with a final reaction volume of 50 μl. Nonspecific signals and contamination were crosschecked using no reverse transcriptase and no template controls. PCR efficiencies of the reactions were >95%. The CFX Manager Software was used to calculate relative expression, whereas standard errors were calculated as in Livak and Schmittgen (2001).

2.4 Image processing

The individual frames of cells were analyzed as follows. First, the cells in the images were segmented using an automatic method (Chowdhury *et al.*, 2013). Next, the each cell intensity is fit to a surface, which is a quadratic polynomial of the distance from the cell border, in least-deviations sense, which is subtracted to obtain the foreground intensity. The foreground intensity is fit with a set of Gaussian surfaces, in least-deviations sense, with decreasing heights until the heights are in the 99% confidence interval of the background noise (estimated assuming a normal distribution and using median absolute deviation). The Gaussians are taken to represent spots, the volume under each representing the total spot intensity. Meanwhile, the volume under the whole foreground surface is taken to represent the total cell intensity. The time series analysis was performed as described in Kandhavelu *et al.* (2012b), from segmentation to spot intensity calculation and RNA estimation. The new procedure was found to perform similar but had lower noise in spot intensities than the method from Kandhavelu *et al.* (2012b).

3 ALGORITHM

3.1 Model of RNA spot intensities

The target RNA contains either 96 or 48 bs (earlier in the text) for MS2d-GFP molecules (Golding *et al.*, 2005). Not necessarily are all the binding sites occupied by GFPs at all moments, but it is reasonable to assume that a large number of the binding sites are occupied, as the MS2-GFP is highly abundant in the cells, and the spots are easily visible at almost all time. The observed intensity can also vary because of other reasons, such as variations of the molecule locations with respect to the focal plane.

If the amounts of light detected from a single GFP are independent and identically distributed, and the binding of MS2-GFP molecules are independent, occurring with a constant probability, the amount of light detected from a single-tagged

RNA should follow a binomially weighted mixture of sums of the amounts of lights detected from single MS2-GFPs. Regardless, if this is strictly true, given a large number of independent sources of light, the light detected from a single-tagged RNA can be well approximated by a normal distribution (central limit theorem), if the signal-to-noise ratio is low. If the signal-to-noise ratio was high, it would be possible to estimate the GFP numbers instead.

Letting the light detected from a single-tagged RNA to follow normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean of μ and variance of σ^2, the light emitted by k-tagged RNAs will be distributed according to $\mathcal{N}(k\,\mu, k\,\sigma^2)$, if the light emitted by the tagged RNAs are independent of one another. If the probability for finding a cluster of k-tagged RNAs is given by α_k, the mixture density takes the form as follows:

$$f_{\mathcal{M}}(x \mid \mu, \sigma^2, \alpha_1, \cdots, \alpha_\infty) = \sum_{k=1}^{\infty} \alpha_k f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2) \tag{1}$$

$$= \sum_{k=1}^{\infty} \frac{\alpha_k}{\sqrt{2\pi\, k\, \sigma^2}} \exp\left(-\frac{(x - k\,\mu)^2}{2\,k\,\sigma^2}\right) \tag{2}$$

where $f_{\mathcal{N}}(x \mid \mu, \sigma^2)$ is the density of a normal distribution with a mean of μ and variance of σ^2.

We do not wish to impose any constraints (model) on the distribution defined by α_k, which represents the RNA distribution, as this distribution is the subject of the study.

3.2 Parameter estimation

The form of the density of Equation (1) makes finding closed-form estimates for the parameters hard. We solve the problem by applying expectation maximization (EM) algorithm (Dempster *et al*., 1977). The EM algorithm iteratively estimates new parameters $\boldsymbol{\theta}'$ using the 'incomplete' (observed) data $\boldsymbol{y}$, the 'complete' data $\boldsymbol{x}$ and the current parameter estimates $\boldsymbol{\theta}$ by maximizing:

$$Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \mathbb{E}[\log f(\boldsymbol{x} \mid \boldsymbol{\theta}') \mid \boldsymbol{y}, \boldsymbol{\theta}] \tag{3}$$

where $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is the complete data density.

We denote the parameters by $\boldsymbol{\theta} \doteq (\mu, \sigma^2, \alpha_1, \cdots, \alpha_N)$. The log-likelihood function for $\boldsymbol{\theta}$ given the intensity observations $\boldsymbol{y} \doteq (y_1, \cdots, y_M)$ and the RNA numbers $\boldsymbol{k} \doteq (k_1, \cdots, k_M)$ is as follows:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{M} \log\left\{ \alpha_{k_i} f_{\mathcal{N}}(y_i \mid k_i\,\mu, k_i\,\sigma^2) \right\} \tag{4}$$

$$= \sum_{i=1}^{M} \log\alpha_{k_i} - \frac{1}{2}\log(2\pi k_i) - \log\sigma - \frac{(y_i - k_i\,\mu)^2}{2\,k_i\,\sigma^2} \tag{5}$$

where $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{k})$ is the complete data. The distribution of the missing parameter K_i under the parameters $\boldsymbol{\theta}$ is given by the following equation:

$$w_{k_i} \doteq \mathbb{P}[K_i = k_i \mid \boldsymbol{y}, \boldsymbol{\theta}] = \frac{\alpha_{k_i} f_{\mathcal{N}}(y_i \mid k_i\,\mu, k_i\,\sigma^2)}{\sum_{k'=1}^{N} \alpha_{k'} f_{\mathcal{N}}(y_i \mid k'\,\mu, k'\,\sigma^2)} \tag{6}$$

which yields the following form for $Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$:

$$Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \sum_{i=1}^{M} \sum_{k_i=1}^{N} w_{k_i}\, \ell(\boldsymbol{\theta}' \mid (y_i, k_i)) \tag{7}$$

The parameters $\boldsymbol{\theta}'$ maximizing Equation (7) can be found by finding the roots of the partial derivatives and verifying that the obtained point is a global maximum. The estimators are as follows:

$$\hat{\alpha}'_k = \frac{1}{M}\sum_{i=1}^{M} w_{k_i} \tag{8}$$

$$\hat{\mu}' = \left(\sum_{k=1}^{N} k\,\hat{\alpha}'_k\right)^{-1} \frac{1}{M}\sum_{i=1}^{M} y_i \tag{9}$$

$$\hat{\sigma}'^2 = \left(\sum_{k=1}^{N} k\,\hat{\alpha}'_k\right)^{-1} \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{N} w_{k_i}\,(y_i - k\,\mu)^2 \tag{10}$$

where the variance estimator involves the true parameter μ. If the estimator $\hat{\mu}'$ is substituted for μ, Bessel's correction should be applied (or the variance will be underestimated on average).

Lastly, we note that in our case the EM iteration is guaranteed to converge to a maximum of the likelihood function (Wu, 1983). However, this maximum is not guaranteed to be the global maximum (Wu, 1983). For this reason, it is required that either the initial parameter values for the EM algorithm are close to the optimal values or multiple initializations are used.

An alternative iterative algorithm with reduced complexity can be derived by assuming some values of k_i estimating the parameters and by assigning new k_i by classifying the data under new estimate of the parameters. This algorithm is similar to the k-means clustering algorithm and is referred here as 'hard' EM, as opposed to the previous being 'soft' EM. Assuming k_i, the parameter estimates become (by letting $w_{k_i} = \mathbb{I}\{k_i = k\}$):

$$\hat{\alpha}'_k = \frac{1}{M}\sum_{i=0}^{M} \mathbb{I}\{k_i = k\} \tag{11}$$

$$\hat{\sigma}'^2 = \left(\sum_{k=1}^{N} k\,\hat{\alpha}'_k\right)^{-1} \frac{1}{M}\sum_{i=0}^{M} (y_i - k_i\,\mu)^2 \tag{12}$$

and for $\hat{\mu}'$ as in Equation (9). In these, $\mathbb{I}\{\cdot\}$ is the indicator function (unity if the condition is true, and zero otherwise), i.e. the estimator $\hat{\alpha}'_k$ is the fraction of items with k_i equal to k. With these, only $\mathcal{O}(M + N)$ work is required per iteration, instead of $\mathcal{O}(MN)$ of the soft EM algorithm, which is significant for large N.

Regardless of the algorithm used, an initial parameter estimate is required. We found the following scheme to be appropriate: (i) sort the observed data y_i and partition it into N bins, (ii) assign $k_i = j$ if y_i is in the jth bin and (iii) estimate initial parameters using Equations (9), (11) and (12). A good partitioning depends on the true values of α_k. We found that equidistant partitioning in y_i is simple, parameter-free and appeared to yield results equivalent to more complicated schemes (e.g. multiclass Otsu's method). It is noted that hard boundaries (as in the above scheme) cause the variance to be initially underestimated, if the overlapping of the clusters is large.

The parameter N can be found by finding the parameter estimates for several values of N and by selecting the model with least order that does fit significantly better than the lower-order models. To determine the significance, we use a likelihood ratio test, where the log-likelihoods ℓ_a and ℓ_b of models of orders a and b are obtained, respectively, and the statistic $-2(\ell_a - \ell_b)$ is computed. For $M \to \infty$, this statistic follows a χ^2 distribution, with $b-a$ degrees of freedom (Wilks, 1938), from which a P-value can be computed. If the P-value is smaller than a given significance level, the higher-order model should be favored over the lower-order one. We note that this procedure rarely selects a model of too high order (as determined by the significance level), but for small sample sizes it might lack evidence to select the appropriate high-order model.

3.3 Spot classification

Next, we construct a classifier that is used to estimate the number of RNAs in a cluster based on the intensity of the cluster. We use maximum a posteriori decision rule, i.e a class k, which is the most probable, is to be associated with an intensity value x:

$$C(x) \doteq \arg\max_k \mathbb{P}[K = k \mid x] \tag{13}$$

$$= \arg\max_k \alpha_k f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2) \tag{14}$$

where $\mathbb{P}[K = k \mid x]$ represents the posterior probability, and α_k the priors and $f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2)$ the likelihood functions of each class (the equality owing to the Bayes rule).

The classification can be performed by evaluating the term to be maximized for each k. Alternatively, a range of intensity values can be associated with each k. Possible decision boundaries can be obtained from the equation:

$$\alpha_a f_{\mathcal{N}}(x^* \mid a\,\mu, a\,\sigma^2) = \alpha_b f_{\mathcal{N}}(x^* \mid b\,\mu, b\,\sigma^2) \tag{15}$$

$$\Rightarrow x^* = \pm\sqrt{a\,b\,(\mu^2 + \sigma^2 \log(a^{-1}\,b\,\alpha_a^2\,\alpha_b^{-2}))} \tag{16}$$

If the decision boundaries x^* do not exist, the density of the higher order envelopes the density of the lower-order one. If so, the lower-order class must not be associated with any intensity. Alternatively, even though the decision boundaries exist, it might be that a lower-order density is enveloped by multiple higher-order ones. A procedure starting from the highest-order density and proceeding to the lowest can determine the decision boundaries in $\mathcal{O}(N)$ time, which enables the lower complexity of the hard EM algorithm. The decision boundaries are symmetric around zero, so the classification can be performed on $|x|$, rather than x.

For the purposes of evaluating the classifier performance, the expected accuracy (ACC) can be computed:

$$\mathbb{E}[\mathbb{I}\{C(X) = K\}] \tag{17}$$

$$= \sum_{k=1}^{N} \alpha_k \int_{[x_k^-, x_k^+] \cup [-x_k^+, -x_k^-]} f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2)\,\delta x \tag{18}$$

where x_k^- and x_k^+ are such that $\forall\, x, 0 \le x_k^- \le x < x_k^+ : C(x) = k$. The integral does not have a closed form solution, but it is the Gauss error function, and can be evaluated numerically. Lastly, we note that this quantity applies asymptotically if the model is true and the estimated parameters are correct. Nevertheless, it is likely useful to evaluate how hard the estimation problem is.

3.4 Means of comparison with the previous method

The previous method of RNA quantification from the spot intensity histogram, here called rounding method, relied on manual inspection of the intensity distribution (Golding *et al*., 2005). Namely, the location of the first peak in the distribution of intensities is selected by an expert, after which the intensities are divided by this value to obtain the RNA numbers in each spot and cell (Golding *et al*., 2005). The discretization is achieved by rounding, which can result in suboptimal choice of thresholds for the classifier accuracy.

Given our model of spot intensities, the expected accuracy of the rounding classifier can be computed using Equation (17) with $x_k^{\pm} = (k \pm \frac{1}{2})\,\Delta$ (with the exception that $x_1^- = 0$ and $x_N^+ = \infty$), where Δ is the location of the first peak. For comparison, we find Δ such that the accuracy is maximized. This classifier is not realizable, as the true parameters must be known, but it serves as the upper limit of performance of the classifier. Alternatively, it is possible to use the parameter estimation procedure proposed with the classification of the rounding method, which has the advantage that finding Δ can be automated. However, on average, such an automated method cannot perform better than the method proposed.

4 RESULTS

4.1 Estimating the number of MS2-GFP-tagged RNAs

We first used our method to estimate the number of MS2-GFP-tagged RNA molecules in live *E.coli* cells. In one case, cells contained an RNA coding for 96 bs for MS2-GFP (Golding *et al*., 2005). In the other case, cells contained a different construct, with only 48 bs. In both cases, cells were induced with 1 mM of IPTG and 0.1% of arabinose, and images were taken 60 min after induction. In theory, we expect the intensity of tagged RNAs to be halved in the second set of measurements.

From both sets of images, we extracted the total pixel intensity of each cell and of each spot. This procedure yielded 269 and 155 samples of spot and cell intensities, respectively, from cells with the 96 bs construct, and 443 and 242 from cells with the 48 bs construct.

For each construct, we assumed that all tagged RNAs exhibit the same fluorescence level when measured from either spot or cell intensities and can be represented by a distribution with the same mean and variance (μ and σ^2). Such mean and variance only differ between the two constructs. With this constraint, one can use both datasets (cell and spot intensities), to jointly estimate the parameters of the model, for each construct. For this, we modified the estimator to account the joint estimation of the two sets of data and estimated the parameters in each case. The distribution of measured intensities along with the estimated distributions is shown in Figure 1, and the values of the parameter estimates are given in Table 1.

First, one would expect the mean of intensities detected from the 48 bs RNAs to be half of the one detected from the 96 bs

RNAs. However, from Table 1, the estimated ratio of their means of intensities is 0.83, rather than one half. One likely explanation for this is that the number of functional binding sites in the two RNA constructs differs from the intended numbers (particularly in the case of the longer construct). Meanwhile, the ratio between the variances is expected to be similar to the ratio between the means. Instead, it equals 0.69, which deviates from the measured value, likely due to deviations in noise levels arising from non-linear changes in intensities with the number of binding sites. Nevertheless, we kept the estimated parameters values (Table 1), rather than imposing the constraint on the values, as they allow a significantly better fit than would be achieved by constraining the ratios to one half, as determined by a likelihood-ratio test ($P < 2.1 \times 10^{-4}$).

Because the induction level of the target gene is the same in the two cases (96 and 48 bs constructs), one also expects similar RNA-per-spot (first and third row of Table 1) and RNA-per-cell distributions (second and fourth row) in both cases. The obtained values of $\hat{\boldsymbol{\alpha}}$ suggest that this expectation appears to be correct. In agreement, we found no evidence that the parameters in Table 1 fit significantly better than in a case where each of the RNA-per-spot and RNA-per-cell distributions are constrained to be equivalent, as determined by a likelihood ratio test ($P > 0.54$).

Next, we used the parameters obtained from the spot intensities and cell intensities to estimate the number of RNAs in each cell. Results are shown in Table 2. We note that the procedure of estimating the RNA numbers using the classification of intensities of each spot is expected to yield better results than estimating the RNA numbers from the total intensities of each cell (cf. accuracy in Table 2), as the problem is easier and the sample size is larger, but it requires an accurate detection of the spots inside the cells.

As expected (and necessary if the methods are appropriate), the RNA statistics (Table 2) are similar in the four cases (even though classifying spots are expected to yield better results than classifying cells). Also, the results are in agreement with previous measurements using the RNA target for 96 bs (Kandhavelu *et al*., 2012b) in terms of mean and variance of RNA numbers.

4.2 Comparison with a time series method

To validate our results, we compared our method with a previously introduced method (Kandhavelu *et al*., 2012b) to extract RNA statistics from time series data (Kandhavelu *et al*., 2012a, b; Muthukrishnan *et al*., 2012). Unlike our method, this method uses temporal information, i.e. time series of intensities in the cells, which allows better accuracy of RNA counting, but makes it unsuitable for analysis of individual frames of cell populations.

We collected images taken for 2 h, separated by 1 min intervals, of cells subject to the same media and induction as in the previous cases. Then, we made use of the method from Kandhavelu *et al*. (2012b) to extract the RNA statistics at ~60 min. The results are shown in Table 3. Visibly, these results are similar to those obtained by our method (cf. Table 2).

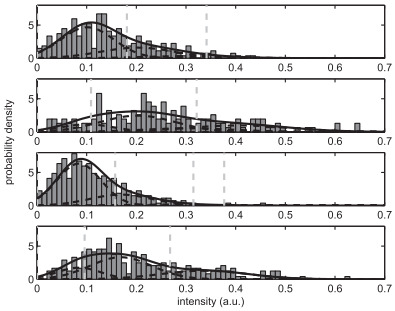

Fig. 1. Distribution of intensities from MS2-GFP-tagged RNA measurements. Panels from top to bottom: spots (96), cells (96), spots (48) and cells (48). The gray bars represent the measured intensity histograms, the solid black lines the estimated distributions, the dashed black lines their components and the dashed gray lines the decision boundaries

Table 1. Estimated model parameters from the MS2-GFP-tagged RNA measurements

Case	M	$\hat{N}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\boldsymbol{\alpha}}$
Spots (96)	269	4	0.101	0.053	(062, 0.30, 0.07, 000)
Cells (96)	155	4	0.101	0.053	(015, 0.46, 0.12, 026)
Spots (48)	443	4	0.084	0.044	(070, 0.26, 0.03, 001)
Cells (48)	242	4	0.084	0.044	(019, 0.52, 0.00, 029)

Note: The table shows the number of samples M, the estimated model order $\hat{N}$ (Section 2), the estimated parameters mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of the intensity of one RNA and the vector of probabilities $\hat{\boldsymbol{\alpha}}$. The value of $\hat{\alpha}_k$ is the estimated probability that one has k RNA molecules in a spot or cell (depending on the case).

Table 2. Estimated distribution of RNAs per cell from the MS2-GFP-tagged RNA measurements

Symbol	Case	M'	μ_r	σ_r^2	ACC
A	Spots (96)	182	1.97	1.55	0.78
B	Cells (96)	182	2.07	1.68	0.68
C	Spots (48)	270	2.02	1.46	0.82
D	Cells (48)	270	2.10	1.48	0.77

Note: The table shows number of cells M', mean μ_r and variance σ_r^2 of RNA numbers per cell and the expected accuracy (ACC).

To verify the agreement between the results using the two methods, we performed two-sample Kolmogorov–Smirnov permutation tests with the null hypothesis of the RNA numbers extracted from different methods and/or cases (A through F in Tables 2 and 3) are generated by an equal distribution, which assesses if the distributions are significantly different. All tests were done with 10^6 permutations, and the results are shown in Table 4.

From Table 4, we find that the results obtained from the images of cell populations using the spot intensities provided similar results to those extracted from time series (i.e. the null hypothesis cannot be rejected when comparing cases A, C, E and F). This indicates that our method performs consistently with the time series method. The same does not hold for the results extracted from the cell populations using cell intensities in all cases, where statistical differences are detected with higher sample sizes (B or D versus C or E). This further suggests that the spot intensities should be used in favor of the cell intensities for more accurate quantification of the RNA numbers. Generally, and confirming the results of the previous section, we found no evidence that the results extracted from the 96 and 48 bs constructs are statistically different.

4.3 Comparison with qPCR measurements

We compared the fold-change in RNA numbers estimated using our method with those obtained by qPCR. For this, we used the 96 bs construct with induction levels of 1 mM (Table 2) and of 0 mM of IPTG (not shown), the former resulting in a higher expression rate. The estimated mean RNA numbers in the case of 0 mM of IPTG using our method were 0.68 and 0.63, using the spot and cell intensities, respectively. These result in expression ratios of 0.345 and 0.303. The expression ratio obtained by qPCR is 0.305 with a standard deviation of 0.024. Both numbers estimated using our method are within the 90% confidence interval of the qPCR measurement, indicating a strong agreement.

Table 3. RNA statistics estimated using time series method (Kandhavelu *et al.*, 2012b) at 60 min after induction

Symbol	Case	M'	μ_r	σ_r^2
E	Cells (96), time series	252	2.09	1.51
F	Cells (48), time series	107	2.02	1.41

Note: The table shows number of cells M' and the mean μ_r and variance σ_r^2 of RNA numbers per cell.

Table 4. Comparison between the RNA distributions extracted different methods and/or data

Symbol	B	C	D	E	F
A	0.027	0.632	0.022	0.696	0.952
B	–	0.00991	0.547	0.00457	0.022
C	–	–	0.00911	0.501	0.872
D	–	–	–	0.00375	0.029
E	–	–	–	–	0.913

Note: *P*-values of the Kolmogorov-Smirnov permutation test under the null hypothesis that a pair of samples of RNA numbers from different cases (A through F) come from the same distribution. A low *P*-value (i.e. less than 0:01) indicates that the RNA distributions are likely inequal.

4.4 Applying the method on simulated data

Lastly, we assessed the performance of the parameter estimation and classification using data from Monte Carlo (MC) simulations of our model [Equation (1)] so that the ground truth is known.

First, we computed the expected accuracy of the classification procedure for different parameters values, which assesses the asymptotic performance of the classifier if the parameters are well estimated. This is shown in Figure 2. The accuracy can be good for either low-mean levels, where the problem is simpler, as most of time the number of RNAs is equal to unity, or for low-noise levels, as in this case the distribution consists of distinct peaks. More importantly, for typical RNA levels in *E.coli* [i.e. 1–10 (Bernstein *et al.*, 2002)], the accuracy is >0.8 for noise levels < 0.25 (which are in agreement with our results in Table 1). For high RNA levels and/or high noise levels, the accuracy deteriorates, as the distribution no longer exhibits distinct peaks.

We found that the expected accuracy is similar with slightly non-Poissonian data (not shown) but generally better with sub-Possonian and worse with super-Poissonian. For typical RNA levels and low-noise levels (<0.1), the accuracy remains >0.9 even for geometric distributed RNA numbers. The rounding method has comparable performance only for limited noise levels, and its performance is more sensitive to the RNA distribution.

With a finite sample, the parameter estimates are not necessarily correct, which causes errors in the classification. We tested the method for various samples sizes, and the results for noise level of $\sigma\,\mu^{-1} = 0.25$ and Poissonian RNA numbers with a mean of $\lambda = 2.5$ are shown in the top panel of Figure 3. For small samples (e.g. 10), the performance is not comparable with the asymptotic performance, whereas for sample sizes $>10^3$ the mean accuracy exceeds the theoretical accuracy of the rounding method (0.8049).

We also used noise level $\sigma\,\mu^{-1} = 0.5$ and a bimodal RNA distribution of $\alpha = (030, 0.01, 0.01, 0.66, 0.01, 001)$ (these are observed e.g. in the case of genes integrated into circuits such

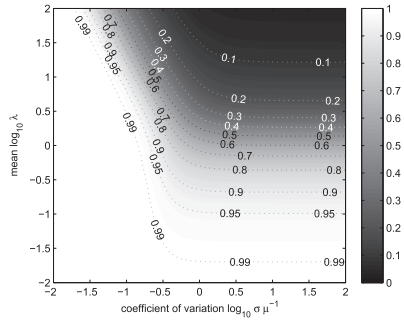

Fig. 2. Expected accuracy of the classification problem for Poisson-distributed RNA numbers. Surface plot of the expected accuracy of the classifier for Poisson-distributed RNA numbers, as a function of noise $\sigma\,\mu^{-1}$ (coefficient of variation) and RNA mean level λ. Light shades of gray represent high accuracy, whereas dark shades represent low

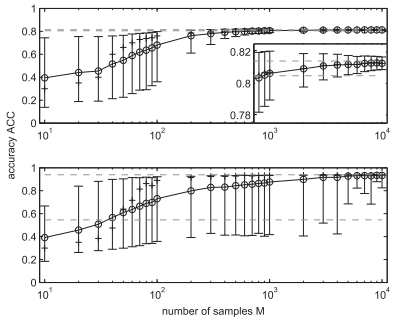

Fig. 3. Distributions of accuracies for various samples sizes M obtained using MC simulations. Top panel: Poisson distributed α, with $\lambda = 2.5$ and $\sigma\,\mu^{-1} = 0.25$. Bottom panel: bimodal distribution of α and $\sigma\,\mu^{-1} = 0.5$. Circles represent means, pluses medians and whiskers upper and lower standard deviations from the mean. The dashed lines are the expected accuracy for our method (higher value) and for the rounding method (lower value)

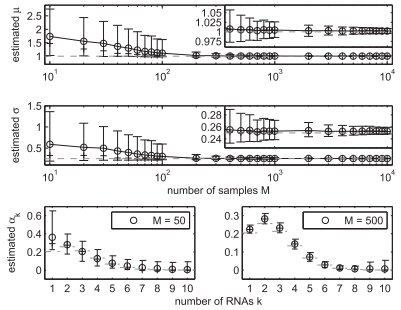

Fig. 4. Distributions of parameter estimates for various sample sizes M obtained using MC simulations. Estimated parameters $\hat{\mu}$ (top panel, true value $\mu = 1$) and $\hat{\sigma}$ (middle panel, true value $\sigma = 0.25$) for various samples sizes. Also shown is the distribution of estimated parameters $\hat{\alpha}_k$, for $M = 50$ (bottom left panel) and $M = 500$ (bottom right). Circles represent means, pluses represent medians and whiskers represent upper and lower standard deviations. The dashed lines are true parameter values

as a toggle switch), shown in the bottom panel of Figure 3. The expected accuracies are 0.9394 and 0.5457, suggesting that our method is appropriate, but the rounding method is not.

Lastly, we note that our method is likely biased for finite samples. We found that for sample sizes <100, it generally overestimates the mean and the variance, which primarily results from underestimation of the order due to lack of evidence for selecting a high–order model with small samples. Regardless, even if the order was known, the maximum likelihood estimator is likely biased. However, these effects are negligible for larger sample sizes ($>10^2$), and e.g. the standard deviation of the parameter estimates exceeds the bias (Fig. 4).

5 DISCUSSION

We have presented a fully automatic method of quantification of RNA numbers from the intensities of the either fluorescent spots or cells. The method consists of a numerical maximum likelihood parameter estimation step (with one of the two proposed methods) followed by a maximum a posteriori classification.

We showed that the method proposed has several advantages. First, by being automated, it will allow an objective comparison of results from independent measurements. Second, our method is expected to have better accuracy than the previous method (Golding *et al*., 2005), when the distribution of RNAs is non-uniform and/or, when the measurement noise is high. When the distribution is uniform and/or the noise-level is low, the solution converges to that of the previous method. Third, our method allows the estimation of its own accuracy. The theoretical analysis indicates that the finite sample biases of the maximum likelihood estimators are negligible and that the method is expected to perform well (~80% accuracy or above) in a typical setting, when the sample size is few hundred samples or more, which is typical for a single-cell study.

The method can be applied on various intensity distributions. We demonstrated its applicability on both the spot intensities and cell intensities extracted from live *E.coli* cells. The choice of the input is two-fold: spot intensities result in more accurate classification, but the detection of spots is required in addition to cell segmentation.

In theory, the method is applicable to any fluorescent or fluorescence-tagged molecules present in low-copy numbers, such as low-expression level fluorescent proteins. However, a proper counting requires a certain degree of separation between brightness levels (not much smaller than the one between e.g. the RNAs with 48 bs). In this regard, our tests showed that the method fails if the data are too noisy or the degree of clustering of the spots is too high. For example, we tested the method on confocal microscopy measurements of tsr-venus proteins coded in *E.coli*, but the signal-to-noise ratio was found to be too low, except for rare cases, where all cells would have one or two proteins. Overall, we expect that, as the methods of fluorescent tagging and microscope improve, our method will become more widely applicable, as it is automatic and allows comparing data from different sources, which is currently not possible.

Funding: Tampere City Science Foundation [to A.H.]; Academy of Finland [257603 to A.S.R.]; Tekes [1386/31/2012 to A.S.R.]; and Fundacao para a Ciencia e Tecnologia [PTDC/BBB-MET/1084/2012 to A.S.R.].

Conflict of Interest: none declared.

REFERENCES

Bernstein,J.A. *et al.* (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl Acad. Sci. USA*, **99**, 9697–9702.

Chowdhury,S. *et al.* (2013) Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE). *BMC Bioinformatics*, **14**, S8.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.*, **39**, 1–38.

Elowitz,M.B. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Golding,I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Golding,I. and Cox,E.C. (2004) RNA dynamics in live *Escherichia coli* cells. *Proc. Natl Acad. Sci. USA*, **101**, 11310–11315.

Kaern,M. *et al.* (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, **6**, 451–464.

Kandhavelu,M. *et al.* (2012a) Regulation of mean and noise of the *in vivo* kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Lett.*, **586**, 3870–3875.

Kandhavelu,M. *et al.* (2012b) Single-molecule dynamics of transcription of the lar promoter. *Phys. Biol.*, **9**, 026004.

Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods*, **25**, 402–408.

Lloyd-Price,J. *et al.* (2012) Asymmetric disposal of individual protein aggregates in *Escherichia coli*, one aggregate at a time. *J. Bacteriol.*, **194**, 1747–1752.

Montero Llopis,P. *et al.* (2010) Spatial organization of the flow of genetic information in bacteria. *Nature*, **466**, 77–81.

Muthukrishnan,A.B. *et al.* (2012) Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucleic Acids Res.*, **40**, 8472–8483.

Ozbudak,E.M. *et al.* (2002) Regulation of noise in the expression of a single gene. *Nat. Genet.*, **31**, 69–73.

Peabody,D.S. (1993) The RNA binding site of bacteriophage MS2 coat protein. *EMBO J.*, **12**, 595–600.

Raj,A. *et al.* (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–879.

Taniguchi,Y. *et al.* (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.

Wilks,S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Satist.*, **9**, 60–62.

Wu,C.F.J. (1983) On the convergence properties of the EM algorithm. *Ann. Stat.*, **11**, 95–103.

Yu,J. *et al.* (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.

Publication II

Hakkinen, A. and Ribeiro, A. S., "Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data," *Bioinformatics*, vol. 31, no. 1, pp. 69–75, 2015.

BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 69–75
doi:10.1093/bioinformatics/btu592

Gene expression

Advance Access publication September 3, 2014

Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data

Antti Häkkinen and Andre S. Ribeiro*

Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, P.O. box 553, 33101 Tampere, Finland

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MS2-GFP-tagging of RNA is currently the only method to measure intervals between consecutive transcription events in live cells. For this, new transcripts must be accurately detected from intensity time traces.

Results: We present a novel method for automatically estimating RNA numbers and production intervals from temporal data of cell fluorescence intensities that reduces uncertainty by exploiting temporal information. We also derive a robust variant, more resistant to outliers caused e.g. by RNAs moving out of focus. Using Monte Carlo simulations, we show that the quantification of RNA numbers and production intervals is generally improved compared with previous methods. Finally, we analyze data from live *Escherichia coli* and show statistically significant differences to previous methods. The new methods can be used to quantify numbers and production intervals of any fluorescent probes, which are present in low copy numbers, are brighter than the cell background and degrade slowly.

Availability: Source code is available under Mozilla Public License at http://www.cs.tut.fi/%7ehakkin22/jumpdet/.

Contact: andre.ribeiro@tut.fi

Received on June 3, 2014; revised on July 14, 2014; accepted on August 27, 2014

1 INTRODUCTION

Transcription, translation and degradation of RNA and proteins are stochastic processes (Elowitz *et al.*, 2002; Kaern *et al.*, 2005; Ozbudak *et al.*, 2002; Taniguchi *et al.*, 2010; Yu *et al.*, 2006). Their stochasticity results in phenotypic variability in monoclonal cell populations (Elowitz *et al.*, 2002; Pedraza and van Oudenaarden, 2005) and temporal phenotypic changes in cells (Golding *et al.*, 2005; Yu *et al.*, 2006). As such, to better understand phenotypic diversity, the mechanisms regulating RNA and protein numbers must be understood.

Presently, the only technique for quantifying RNA numbers and transcription dynamics in live *Escherichia coli* cells over time consists of tagging target RNA molecules with an array of fluorescent MS2-GFP proteins (Golding and Cox, 2004). Using this technique, the target RNA can be visualized as a bright spot (see e.g. Fig. 6) shortly after production (Golding and Cox, 2004; Golding *et al.*, 2005). This method allows studying transcription in the absence of several sources of stochasticity (Kandhavelu *et al.*, 2012a), such as RNA degradation (Taniguchi *et al.*, 2010) or dilution by cell division (Huh and Paulsson, 2011).

The quantification of transcription dynamics from fluorescence microscopy images requires estimation of the RNA numbers or the RNA production times using statistics extracted from microscopy images, such as temporal intensity signals. Some methods have been proposed for determining RNA numbers using cell and/or fluorescence spot intensities, either using manually assisted (Golding *et al.*, 2005) or automatic (Hakkinen *et al.*, 2014) techniques. However, these methods were designed to extract stationary RNA distributions from cell populations, and as such, they neglect any temporal information in the data. Recently, we introduced a method (Kandhavelu *et al.*, 2012b) to use such information to extract time intervals between consecutive RNA productions in individual cells. This method uses a piecewise-constant monotonic least-squares (LSQ) fit with an *F*-test to select the model order, after which a jump in the model curve is taken to correspond to the production of a target RNA. Unfortunately, this method does not extract the absolute RNA numbers.

Here, we propose a new method for automatic quantification of RNA numbers and RNA production intervals from intensity time series extracted from cells. Specifically, we consider two variants: one that uses LSQ costs, which can be derived using the central limit theorem, and one that uses least-deviations (LD) costs, which is a robust variant of the former. In particular, the latter was designed to counter outliers commonly observed in the intensity data, caused by e.g. RNA molecules moving out of the focal plane, at the cost of reduced accuracy. The new method was designed to exploit the temporal information in the data for improved accuracy, to not require post- and/or pre-processing for time interval extraction and to be free of any regularization in temporal domain to allow more accurate quantification of time intervals.

First, we present Monte Carlo simulations to demonstrate that the accuracy of the method is, in general, superior to the existing methods, in estimating both RNA numbers and RNA production intervals. Second, we apply our method and the previous methods on novel data extracted from time-lapse microscopy measurements of live *E.coli* cells expressing MS2-GFP and RNA target to show that, for large number of cells, statistically significant differences in the results can be detected between the new and previous methods, in both RNA numbers and RNA production time intervals.

*To whom correspondence should be addressed.

2 METHODS

2.1 Cells and plasmids

Escherichia coli strain DH5α-PRO was generously provided by I. Golding (University of Illinois) and contains two constructs: a PROTET-K133 medium-copy vector carrying a MS2-GFP reporter, controlled by $P_{LtetO-1}$, and the pIG-BAC single-copy vector coding for mRFP1-MS2-96bs RNA, whose expression is controlled by $P_{lac/ara-1}$ (Golding and Cox, 2004).

2.2 Microscopy

Cells were grown in Miller lysogeny broth (LB) medium supplemented with antibiotics according to the specific plasmids. Cells were grown overnight at 37°C with aeration, diluted into fresh medium and allowed to grow at 37°C until an optical density OD600 of 0.3–0.5 was reached. To attain full induction of the MS2-GFP reporter, cells were incubated with 100 ng/ml of anhydrotetracycline (aTc, from IBA GmbH). In all, 0.1% of L-arabinose (Sigma-Aldrich) and 1 mM of Isopropyl-β-D-thiogalactopyranoside (IPTG, Fermentas) were used to induce the target RNA. Cells were pre-incubated with arabinose at the same time as aTc. IPTG was added 1 h after aTc, and cells were incubated for 5 min.

Microscopy was performed using a Nikon Eclipse (TE-2000-U, Nikon, Tokyo, Japan) inverted confocal laser scanning microscope. Cells were imaged in a thermal chamber set to 37°C. Images were taken 5 min after induction by IPTG, for a duration of 2 h, once per minute. For imaging, a few microliter of culture were placed between a coverslip and a slab of 1% agarose containing LB along with the appropriate concentrations of inducers. When both the reporter and the target RNA are present in the cells, MS2-GFP proteins bind to the target RNA, forming a bright fluorescent spot (Golding *et al.*, 2005). The RNA becomes visible during, or shortly after, elongation (Golding and Cox, 2004).

2.3 Image processing

Cells were detected from the microscope images using a semi-automatic method described in Kandhavelu *et al.* (2012b). First, a mask is manually painted over the area that a cell occupied during the time series. Principal component analysis is then used to obtain the dimensions and orientation of the cells from the fluorescence distribution within each mask. Next, target RNA spots were automatically segmented using kernel density estimation with a Gaussian kernel. Cell background-subtracted spot intensities were then calculated and summed for each cell to produce the total spot intensity within each cell, which was used to quantify RNA numbers.

3 ALGORITHM

3.1 Overview of the method

As each RNA is tagged by a large number (up to 96) of MS2-GFP molecules (Golding *et al.*, 2005), the intensities detected from a single RNA are expected to be well approximated by a normal distribution (central limit theorem). Consequently, sums of intensities of k RNA spots (e.g. in a cell) can be assumed to be normally distributed, with mean and variance k times that of the individual RNA, provided that the components are sufficiently independent (Hakkinen *et al.*, 2014).

Another assumption exploited by our method is that the MS2-GFP-tagged RNA molecules are virtually immortal during a cell lifetime [verified in Muthukrishnan *et al.* (2012)]. Because of this, one can assume that the RNA numbers form a non-decreasing series over time. For that, for example, the data before and after cell division must be treated as separate series.

We propose two variants of our method that use the same strategy but different intensity models. The LSQ variant uses normally distributed errors, as discussed above, but with constant (as opposed to linear) variance. We experimented with the linear variance model, but in the zero-noise limit, it approaches to the constant variance one, and the results were similar. Also, if additional noise sources affecting all the spots in an equal manner (regardless of their RNA numbers) are present, the model should be affine, i.e. somewhere between constant and linear.

The LSQ variant is similar to the method introduced in Kandhavelu *et al.* (2012b), which uses a piecewise-constant monotonic LSQ fit with an F-test to select the model order, after which a jump in the fit curve is taken to correspond to a production of an RNA. However, there the jump sizes are regulated by the means of the F-test (and the monotonicity constraint). Moreover, the F-test also causes regularization in the temporal direction. In our method, regularization is performed via constant jump size and the monotonicity constraint, to avoid regularization in the temporal domain.

Meanwhile, the LD variant was conceived to mitigate the problem that the RNA intensity time series sometimes contain 'holes', caused by, for example, RNA spots moving out of the focal plane (see e.g. the lower panel of Fig. 6). This variant uses the median as the location estimator, making it more robust than LSQ. If no such outliers are present, the LSQ should be preferred, as it is expected to be more accurate.

Regardless of the variant, the method operates as follows:

(1) A curve (without quantization) is fit to the intensity time series. This groups the related samples to reduce uncertainty, extracting the temporal information from the data. If the data are not temporal, this step is effectively a no-operation.

(2) Jump size (and other parameters such as uncertainty) is estimated from the fit pieces. We provide a new set of estimators for this step, but the one proposed in Hakkinen *et al.* (2014) could be used instead.

(3) A quantized curve is fit to the time series, given the parameters, enforcing the quantization to the fit. This is used to provide the RNA numbers.

3.2 Curve fitting

Piecewise-constant curve fitting can be done in polynomial time using a dynamic programming technique. Let $d(x, y)$ be some metric and $(x_i)_{i=1}^{n}$ be some sequence of length n, x_i denoting its ith element. Let $D_{a,b}^{k}$ be the distance between a substring $(x_i)_{i=a}^{b}$ of the input and a k-piece model:

$$D_{a,b}^{k} \doteq \sum_{i=a}^{p_1} d(x_i, \mu_1) + \ldots + \sum_{i=p_{k-1}+1}^{b} d(x_i, \mu_k) \tag{1}$$

which is to be minimized for $a = 1$, $b = n$ and some $k \leq n$:

$$\min_{\substack{p_1,\ldots,p_{k-1},\\ \mu_1,\ldots,\mu_k}} D_{1,n}^{k} = \min_{p_{k-1}} \left(\min_{\substack{p_1,\ldots,p_{k-2},\\ \mu_1,\ldots,\mu_{k-1}}} D_{1,p_{k-1}}^{k-1} + \min_{\mu_k} D_{p_{k-1}+1,n}^{1} \right) \tag{2}$$

which can be computed by memoizing the minima of each $D_{1,b}^{k}$ and $D_{a,b}^{1}$. Given that $D_{1,b}^{k-1}$ have already been minimized, it takes $\mathcal{O}(n^2)$ time to minimize $D_{1,b}^{k}$ for all b. If the minima of $D_{a,b}^{1}$ are computed for all a, b in time $T(n)$, $D_{1,n}^{k}$ can be minimized in

$T(n)+\mathcal{O}(k\,n^2)$ time and $\mathcal{O}(n)$ space. In the process, the solutions for all b and for all orders up to k are obtained. To fit the data with a monotonic function, the solutions violating the monotonicity constraint can be ignored in the minimization process.

In particular cases, a link between the choice of the metric $d(x, y)$ and maximum likelihood estimation can be established. If one assumes some piecewise-constant curve, corrupted by additive independent zero-mean normal distributed errors, a curve fit by the above procedure using squared error metric $d(x, y)=(x-y)^2$ corresponds to the maximum likelihood estimate of the signal. Similarly, there is a link between using absolute error metric $d(x, y)=|x-y|$ and Laplace distributed errors.

We implemented LSQ and LD metrics. The LSQ problem can be trivially implemented in $T(n)=\mathcal{O}(n^2)$ time, by computing running averages of the data. Similarly, the LD problem can be solved by computing running median, and can be implemented in $T(n)=\mathcal{O}(n^2\log n)$ time using priority queues with $\mathcal{O}(\log n)$ update times. In practice, we have been using these methods for up to 10 000 samples on a standard personal computer.

3.3 Jump size estimation

We estimate the jump size using a maximization of approximate likelihood in the low-noise limit. This method is expected to work when the noise-to-signal ratio of the data is reasonably small (see results for examples).

The probability density function of an equiprobable mixture of normal distributions with means at $k\,\mu$ for $k\in\mathbb{Z}$ and variance of σ^2 is as follows:

$$f(x)=\sum_{k=-\infty}^{\infty}\frac{C}{\sqrt{2\pi\,\sigma^2/w}}\exp\left(-\frac{(x-k\,\mu)^2}{2\,\sigma^2/w}\right) \tag{3}$$

where x is the intensity and w is used to scale the uncertainty of x. The constant C is selected such that the density integrates to unity. Constructing the infinite summation as a limit suggests that scaling of $C\propto\mu$ is appropriate. In the limit $\sigma^2\to 0$, the density $f(x)$ can be approximated (disregarding the scale) by the following:

$$g(x)=\frac{\mu}{\sqrt{2\pi\,\sigma^2/w}}\exp\left(-\frac{(x-K(x)\,\mu)^2}{2\,\sigma^2/w}\right) \tag{4}$$

For n independent and identically distributed samples, the approximate likelihood $\prod_{i=1}^{n}g(x_i)$ will be maximized when the squared coefficient of variation σ^2/μ^2 is minimized. This can be found by finding the roots of the partial derivatives of the function and verifying that they are maxima, yielding the following estimators:

$$\hat{\mu}=\frac{\sum_{i=1}^{n}w_i\,x_i^2}{\sum_{i=1}^{n}w_i\,K(x_i)\,x_i} \tag{5}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}w_i\,(x_i-K(x_i)\,\mu)^2 \tag{6}$$

The $K(x)$ that minimizes the approximation error is $K(x)=\lfloor x/\mu\rceil$, where $\lfloor\cdot\rceil$ denotes rounding to the nearest integer. As $K(x_i)$s depend on μ, the problem is not yet solved. However, for each choice of $K(x)$ there is a range of associated μs involved. The problem can be solved by finding the μ and the likelihood for each set of $K(x)$, which is feasible if the values of $K(x)$ are small. Specifically, the solution can be computed incrementally in $\mathcal{O}(k\,n\log n)$ time and $\mathcal{O}(n)$ space using a priority queue with $\mathcal{O}(\log n)$ update times, where k is a bound for $K(x)$.

For large μ, $K(x)\to 0$ and the squared coefficient of variation, $\hat{\sigma}^2/\mu^2$ is around $\frac{1}{n}\sum_{i=1}^{n}w_i\,x_i^2/\mu^2$. On the other hand, for $\mu\to 0$, $x_i-K(x_i)\,\mu$ becomes a uniform random variable in $[-\frac{1}{2},\frac{1}{2})$ and $\hat{\sigma}^2/\mu^2\to\frac{1}{n}\sum_{i=1}^{n}\frac{1}{12}w_i$. Equating the two, an upper bound for the search can be obtained, to avoid the trivial solution of the large μ model. We have not yet devised any better stopping condition than to stop the search when $K(x)$ become sufficiently large [e.g. 100 if one considers typical RNA numbers in *E.coli* (Taniguchi *et al.*, 2010)].

A similar procedure can be applied for Laplace distributed rather than normal distributed data. The estimator for the absolute deviation b is as follows:

$$\hat{b}=\frac{1}{n}\sum_{i=1}^{n}w_i\,|x_i-K(x_i)\,\mu| \tag{7}$$

where $|\cdot|$ denotes the absolute value. The estimator for the location parameter μ is the multiplicative inverse of the $w_i\,x_i$ weighted median of $K(x_i)/x_i$, which is the minimizer of the coefficient of mean deviation $\hat{b}/\mu$.

As the Laplace distribution also has a density that is symmetric about the mode and decreases away from it, the $K(x)$ that minimizes the approximation error is as in the LSQ case. The asymptotic behavior for $\mu\to 0$ is that $\hat{b}/\mu\to\frac{1}{n}\sum_{i=1}^{n}\frac{1}{4}w_i$, and for large μ, the statistic $\hat{b}/\mu$ approaches a value of $\frac{1}{n}\sum_{i=1}^{n}w_i\,|x_i|/\mu$, which can again be used to obtain an upper bound for the search.

3.4 Quantization

The procedure of obtaining a fit curve with quantization, which is the third step in our method, depends on the choice of the error metric. In the case of squared error metric, a fit signal with quantization can be obtained from the fit performed without quantization. This is possible, as for $d(x, y)=(x-y)^2$ and some $Q(\mu)$:

$$\sum_{i=a}^{b}d(x_i,Q(\mu))=\sum_{i=a}^{b}d(x_i,\mu)+d(\mu,Q(\mu)) \tag{8}$$

if μ is a minimizer of $\sum_{i=a}^{b}d(x_i,\mu)$. The minimizer of $d(\mu,Q(\mu))$ is $Q(\mu)=\lfloor\mu/q\rceil\,q$, where q is the chosen quantization level, i.e. in the LSQ case, quantization can be performed by rounding the fit signal to the nearest multiple of the estimated jump size.

The same does not hold for the absolute error metric. For finding the quantized result, we use the curve fitting procedure again with an additional constraint, which imposes the quantization.

4 RESULTS

4.1 Monte Carlo simulations

We performed Monte Carlo simulations with various parameters of a general simple model of appearance of molecules inside cells.

In this model, molecules are generated with some intervals, whose durations are exponentially distributed. For reasons described in the methods, no degradation is modeled. Finally, the molecule numbers are corrupted by adding zero-mean independent and identically normal distributed noise.

First, we generated the true curves using exponentially distributed production intervals with a rate of $(15\,\text{min})^{-1}$ and then corrupted them by adding normally distributed noise with a coefficient of variation of $\sigma\mu^{-1}$ of 0.5, 1 or 2. A total of 100 series were used for analysis in each case, which was performed using the LSQ estimator. Figure 1 shows example time series that were generated by sampling every 10 s for 2 h. For quantifying the performance of the method, we estimated its accuracy, i.e. the proportion of correct RNA estimates in all estimates. The measured accuracies from each simulation were 0.992, 0.965 and 0.880 for $\sigma\mu^{-1}$ of 0.5, 1 and 2, respectively. In this setting, the accuracy remains good even for moderate noise levels, such as for $\sigma\mu^{-1}=2$, despite the fact that the jump size estimator was derived in the zero-noise limit assumption.

Next, in Figure 2, we show the results of testing for sampling frequency of $(1\,\text{min})^{-1}$ with noise level of $\sigma\mu^{-1}=1$ and using both the LSQ and LD estimators. In addition, samples were zeroed independently with a 25% probability to corrupt the data further (bottom case in Fig. 2), so as to test the resistance of the LD estimator to 'holes' (for reference, the LSQ estimator breaks around 1% of zeros in this case). Other parameters were as specified in the previous paragraph. The measured accuracies were 0.834, 0.808 and 0.621, for LSQ and LD without zeroing, and LD with zeroing, respectively. In these cases, the performance is worse, as less data are provided to the method, but the results remain likely useful.

Next we quantified the accuracies for the LSQ method using 1000 independent simulations. We used two sampling settings (10 s and 1 min intervals for a duration of 2 h) and exponential production intervals with a rate of $(15\,\text{min})^{-1}$. For the tests, various noise levels $\sigma\mu^{-1}$ and number of series were used. In addition to the LSQ method, we computed an upper bound for the accuracy of any method that uses only the intensity values and no temporal information–this represents a comparison with a previous method proposed in Hakkinen *et al.* (2014). The results for the 10 s sampling intervals are shown in Figure 3 and for 1 min sampling interval are shown in Figure 4.

For 10 s sampling interval, the accuracy remains high (>0.8) for noise levels up to $\sigma\mu^{-1}=2.5$ for large number of series (> 100), or up to $\sigma\mu^{-1}=2$ for smaller numbers (10). Using >1000 series provides no significant improvement in accuracy. For lower noise levels, the accuracy is generally high, regardless of the number of series. In this setting, the performance of a non-time series (non-ts) method is not comparable, as the noise levels are high and the vast amount of temporal information available is neglected.

The results for 1 min sampling interval exhibit less differences between the two methods in our setting. The accuracy remains high (>0.8) for noise levels up to $\sigma\mu^{-1}$ for large number of series (>100) and up to $\sigma\mu^{-1}=0.75$ for small numbers (10). The figure also shows that around 100 samples are sufficient, and even smaller sample sizes can be used for lower noise levels, in this setting. For high noise levels, both methods are likely to perform poorly, whereas for small noise levels, both methods work well. Importantly, the new method is advantageous for moderate noise levels.

We also varied other parameters, such as the shape of the production interval distribution. Changing the production interval to a more noisy distribution (e.g. gamma distribution with shape parameter <1) results in reduced accuracy, whereas changing it to a lesser noisy distribution results in improved accuracy. Similarly, using the LD classifier for data generated using a model with normal errors results in slightly reduced accuracy when compared with the LSQ method.

Finally, we tested the performance of using our method in estimating the production interval distribution from the results of the fit. For comparison, we also present results using a previous method (Kandhavelu *et al.*, 2012b). The goodness was assessed by computing the Kullback–Leibler divergence of the true distribution from that of the intervals extracted after applying one of the methods—this is the criterion minimized by a

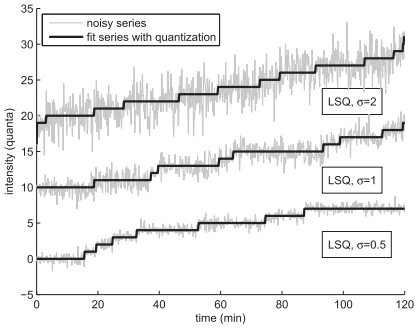

Fig. 1. Example results of the LSQ method with 100 series (first of which is shown) with a duration of 2 h and sampled every 10 s, for various noise levels. The series were generated using a jump size of $\mu = 1$ and exponential intervals with a rate of $(15\,\text{min})^{-1}$

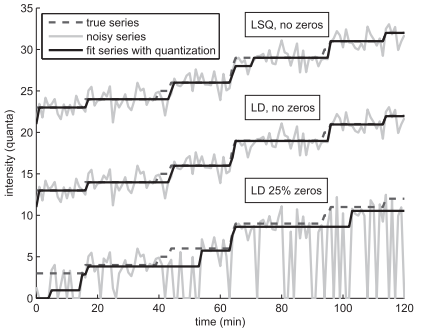

Fig. 2. Different methods tested on 100 series (first of which is shown), each with a duration of 2 h, sampled every 1 min. The series were generated using a jump size of $\mu = 1$, noise level of $\sigma = 1$ and exponential intervals with a rate of $(15\,\text{min})^{-1}$. In the bottom series, a sample is independently zeroed with a probability of 0.25. Most of the time, the true series is covered by the equal fit one

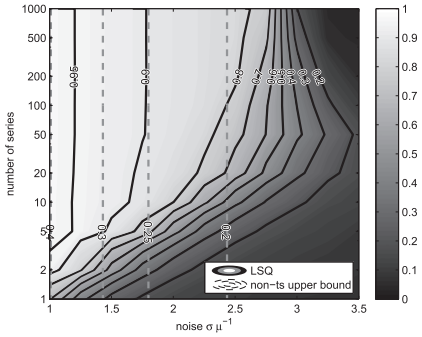

Fig. 3. Mean accuracy of the LSQ method applied to each of 1000 Monte Carlo simulations and the corresponding upper bound for a non-time series type method as a function of the noise $\sigma\mu^{-1}$ and the number of series. The plot was obtained with exponentially distributed production intervals with a rate of $(15\,\text{min})^{-1}$ for a duration of 2 h sampled every 10 s

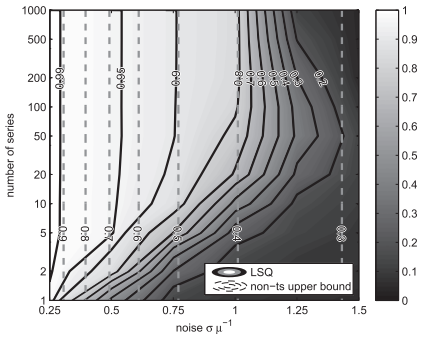

Fig. 4. Mean accuracy of the LSQ method applied to each of 1000 Monte Carlo simulations and the corresponding upper bound for a non-time series type method as a function of the noise $\sigma\mu^{-1}$ and the number of series. The plot was obtained with exponentially distributed production intervals with a rate of $(15\,\text{min})^{-1}$ for a duration of 2 h sampled every 1 min

maximum likelihood estimator. The divergence was computed using a parametric method.

For this, we first varied the shape and mean of the production interval distribution. This was done by using gamma distributed production intervals, with shape $\alpha \in \{0.5, 1, 2\}$, where $\alpha = 1$ yields an exponential distribution, and $\alpha < 1$ and $\alpha > 1$ yield less and more noisy distributions, respectively. The mean duration of the production interval was varied in the range $\beta^{-1} \in [1, 100]$ min. For each simulation, 100 series were generated for the duration of 2 h with sampling intervals of 1 min and noise level of $\sigma\mu^{-1} = 0.5$, and the results were averaged from 1000 simulations. The results for this case are shown in the upper panel of Figure 5. Next, we varied the noise $\sigma\mu^{-1} \in [0.1, 1]$ of the intensities along with the shape of the production interval distribution with a constant mean production rate of

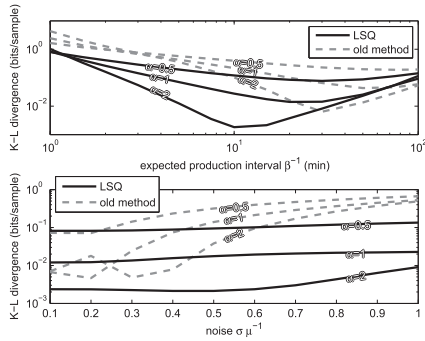

Fig. 5. Kullback–Leibler divergence estimated from 1000 Monte Carlo simulations for estimating the parameters of the interval distribution, as a function of mean production rate β^{-1} (upper panel) and as a function of intensity noise $\sigma\mu^{-1}$ (lower panel). The results were obtained with gamma distributed production intervals with shapes of $\alpha \in \{0.5, 1, 2\}$, for a duration of 2 h sampled every 1 min. Unless otherwise specified, the mean production rate is $\beta^{-1} = 15$ min and the noise is $\sigma\mu^{-1} = 0.5$

$\beta^{-1} = 15$ min. All other parameters were as specified in the previous case. These results are shown in the lower panel of Figure 5.

The results in Figure 5 suggest that, generally, the new method outperforms the previous method. The previous method was found to perform better with specific parameter ranges, particularly for more deterministic ($\alpha = 2$) production intervals, likely owing to tighter regularization. As expected, higher variance in the production interval results in reduced performance in all cases. Similarly, short or long production intervals result in reduced performance, as in the former case, both methods suffer from poor time resolution and in the latter from lack of observed intervals. In the case of low noise $\sigma\mu^{-1}$ in the intensities, both methods have similar performance; however, the performance of the new method is superior for moderate to high noise levels.

4.2 Statistics of tagged RNA numbers and intervals

Finally, we used our method to estimate statistics related to the production of MS2-GFP-tagged RNA molecules in live *E.coli* cells from time-lapse images obtained by confocal microscopy (see Section 2). Each cell contains one lac/ara-1 promoter expressing the target RNA, which consists of 96 binding sites for the MS2-GFP for visualization, preceded by a sequence coding for mRFP (see upper panel of Fig. 6). Target and reporter genes were induced as described in the methods, and images were taken 60 min after induction, sampled every 1 min for a duration of 2 h. The image analysis procedure (see Section 2.3) of three independent experiments performed in the same conditions produced 503 cells with a total of 24 466 intensity samples.

For comparison, we also used the two previous methods used in the previous section. The first, proposed in Hakkinen *et al.* (2014), does not use the temporal information as our method does. Consequently, we expected it to yield less accurate results, as suggested by results on the Monte Carlo simulations shown in the previous section. The other previous method was proposed

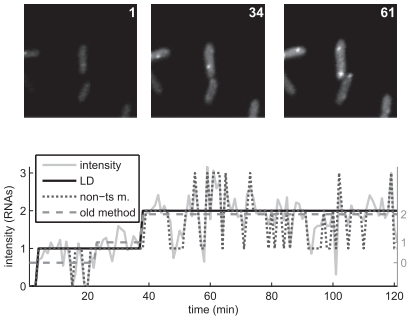

Fig. 6. Upper panel: Regions of confocal microscope images at 1, 34 and 61 min after starting the imaging. Lower panel: example intensity series and fit curves using the LD method and the two previous methods. The curves are shown in the space of RNA numbers estimated by the LD method; corresponding RNA numbers using the old method are shown on the right. The centermost cell of the images in the upper panel corresponds to that shown in the lower

and first used in Kandhavelu *et al.* (2012b). Again, based on the results from the previous section, we expect our method to be slightly more accurate than this method.

An example of a time series of cell fluorescence intensities extracted from the microscopy data is shown in the lower panel of Figure 6. The RNA curves fit using each of the three methods for the example series are also shown. Visibly, the intensity time series contains moderate amount of noise and occasional samples with small values (cf. the bottom curve in Fig. 2). Because of the latter observation, we opted to use the LD variant for the analysis.

The fit curves in the lower panel of Figure 6 exemplify how the three methods operate using different strategies. Without any pre- or post-processing, the non-ts method seems ill-suited. Also, the example shows how the LD method is regularized in intensity space and the previous method in temporal space. Nevertheless, the results from the two methods tend to generally agree with each other—both insert RNA productions when large increases in the intensity series occur.

We compared the values of RNA numbers extracted using the three methods from the data of the experiments. For the purposes of comparison, we extracted the RNA numbers averaged over each cell and each point in time. As these numbers appeared to be similar, we performed a Kolmogorov–Smirnov (K-S) test to assess whether the RNA numbers extracted using either of the previous methods differ from that of our LD method, in a statistical sense. In this test, the null hypothesis is that the two sets of data are generated from an equal distribution, and the alternative hypothesis is that the distributions are unequal. As the data are discrete and possibly correlated over time, the test was performed by permuting the series (rather than individual samples) for a total of 10^6 times. The results are summarized in Table 1.

For the obtained time series of intensities, we also extracted the RNA production intervals using both our LD method and the previous method designed for production interval extraction (Kandhavelu *et al.*, 2012b). The mean and the squared coefficient of variation of the extracted intervals are shown in Table 2.

Table 1. Statistics of RNA numbers

Method	Cells	Samples	Mean	K-S *P*-value
LD	503	24 466	1.43	N/A
Non-ts method	-"-	-"-	1.49	2.0×10^{-5}
Old method	-"-	-"-	1.44	2.8×10^{-4}

Notes: The table lists the number of cells, number of RNA samples, mean RNA numbers and *P*-value of the K-S test when comparing with the LD method.

Table 2. Statistics of time intervals between RNA productions

Method	Intervals	Mean (min)	Squared-CV	K-S *P*-value
LD	373	15.42	0.82	N/A
Old method	344	16.09	0.52	2.1×10^{-4}

Notes: The table lists the number, mean, and squared coefficient of variation (squared-CV) of the extracted intervals, and the *P*-value of the K-S test when comparing to the LD method.

Also, a K-S test was performed to assess the statistical significance of the differences between the interval distributions.

In summary, the three methods extracted similar but statistically distinct RNA numbers and durations between the productions of consecutive RNA molecules from the data. Also, in all cases, the results agree with those reported in Kandhavelu *et al.* (2012a). In particular, while the mean numbers these quantities appear similar, the K-S tests were able to detect significant differences at the level of resolution of the measurements. This, along with the evidence presented in the previous section, suggests that our method offers significant improvements over the methods previously used for such studies. Interestingly, the results using the LD method suggest slightly noisier shape of the distribution than previously reported (Kandhavelu *et al.*, 2012a). This is likely to be of relevance to studies of the mechanisms underlying transcription in live cells. Further, might suggest that our new method produces more accurate results by avoiding the regularization in the temporal domain.

5 DISCUSSION

We have presented two variants of a novel, more accurate method for the automatic quantification of RNA numbers and RNA production intervals from intensity time series extracted from images of live cells expressing fluorescently tagged RNA molecules. The new method exploits the temporal information in the data for improved accuracy, does not require post- and/or pre-processing for time interval extraction and has no regularization in the temporal domain.

One of the proposed variants uses LSQ costs, which can be derived using the central limit theorem. The other uses LD costs, which is a robust variant of the former, and is to be used when there is potential for outliers (e.g. spots transiently leaving the focal plane of the microscope). Meanwhile, the former is preferred when no such corruption is present (e.g. if using multiple slices along the *z*-axis at each time point).

We used Monte Carlo simulations to demonstrate that the accuracy of the new methods is, in general, superior to that of our two previously proposed methods (Hakkinen *et al*., 2014; Kandhavelu *et al*., 2012b), both in estimating the RNA numbers and the RNA production intervals. We also applied the new and the previous methods on novel data from time-lapse images of live *E.coli* cells expressing RNA target for MS2-GFP to show that, if the data contains large number of cells, statistically significant differences in the results can be detected, in both the RNA numbers and RNA production time intervals. In this regard, it should be noted that such 'large' numbers of cells are required to, e.g. compare changes in the dynamics of RNA production by a promoter when under different temperatures or levels of induction (Kandhavelu *et al*., 2012a; Makela *et al*., 2013).

Currently, MS2-GFP-tagging of RNA molecules is the only existing method for detecting RNA molecules, as they are produced in live cells. An accurate quantification of the copy numbers of these tagged RNA molecules and determination of the moments when they first appear are essential, particularly in studies of the dynamics of transcription in live cells [see e.g. Muthukrishnan *et al*. (2012)]. Such measurements are currently being used to assess, for example, the role of induction and repression mechanisms in regulating gene expression dynamics or the effects of environmental factors such as temperature on this dynamics. However, detection and quantification of individual, tagged RNA molecules, and thus our method, are valuable to other endeavors as well. For example, recent uses of the counting of such molecules as they appear in cells include a study of the dynamics of small genetic circuits (Chandraseelan *et al*., 2013) and a study of errors in the partitioning of RNA molecules in cell division (Lloyd-Price *et al*., 2012).

The two variants of the method proposed here should prove valuable in increasing the accuracy of these, as well as of other studies making use of fluorescent molecules, provided that the fluorescent molecules exist in small numbers in each cell, that their fluorescence is significantly above the background fluorescence and that they have a slow degradation rate when compared with dilution caused by cell division.

Funding: Work supported by Jenny and Antti Wihuri Foundation [to A.H.]; Academy of Finland [257603 to A.S.R.]; Tekes [40226/12 to A.S.R.]; and Fundacao para a Ciencia e Tecnologia [PTDC/BBB-MET/1084/2012 to A.S.R.].

Conflict of interest: none declared.

REFERENCES

Chandraseelan,J.G. *et al.* (2013) Temperature dependence of the LacI-TetR-CI repressilator. *Mol. Biosyst*., **9**, 3117–3123.

Elowitz,M.B. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Golding,I. and Cox,E.C. (2004) RNA dynamics in live *Escherichia coli* cells. *Proc. Natl Acad. Sci. USA*, **101**, 11310–11315.

Golding,I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Hakkinen,A. *et al.* (2014) Estimation of fluorescence-tagged RNA numbers from spot intensities. *Bioinformatics*, **30**, 1146–1153.

Huh,D. and Paulsson,J. (2011) Random partitioning of molecules at cell division. *Proc. Natl Acad. Sci. USA*, **108**, 15004–15009.

Kaern,M. *et al.* (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet*., **6**, 451–464.

Kandhavelu,M. *et al.* (2012a) Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Lett*., **586**, 3870–3875.

Kandhavelu,M. *et al.* (2012b) Single-molecule dynamics of transcription of the lar promoter. *Phys. Biol*., **9**, 026004.

Lloyd-Price,J. *et al.* (2012) Probabilistic RNA partitioning generates transient increases in the normalized variance of RNA numbers in synchronized populations of *Escherichia coli*. *Mol. BioSyst*., **8**, 565–571.

Makela,J. *et al.* (2013) *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Res*., **41**, 6544–6552.

Muthukrishnan,A.-B. *et al.* (2012) Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucleic Acids Res*., **40**, 8472–8483.

Ozbudak,E.M. *et al.* (2002) Regulation of noise in the expression of a single gene. *Nat. Genet*., **31**, 69–73.

Pedraza,J.M. and van Oudenaarden,A. (2005) Noise propagation in gene networks. *Science*, **307**, 1965–1969.

Taniguchi,Y. *et al.* (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.

Yu,J. *et al.* (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.

Publication III

Hakkinen, A. and Ribeiro, A. S., "Characterizing rate limiting steps in transcription from RNA production times in live cells," *Bioinformatics*, in press, doi: 10.1093/bioinformatics/btv744, 2015.

Bioinformatics Advance Access published December 31, 2015

Characterizing rate limiting steps in transcription from RNA production times in live cells

Antti Häkkinen [1] and Andre S. Ribeiro [1,*]

[1]Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology, P.O. box 553, 33101, Tampere, Finland

Associate Editor: Dr. Ziv Bar-Joseph

ABSTRACT

Motivation: Single-molecule measurements of live *Escherichia coli* transcription dynamics suggest that this process ranges from sub- to super-Poissonian, depending on the conditions and on the promoter. For its accurate quantification, we propose a model that accommodates all these settings, and statistical methods to estimate the model parameters and to select the relevant components.

Results: The new methodology has improved accuracy and avoids overestimating the transcription rate due to finite measurement time, by exploiting unobserved data and by accounting for the effects of discrete sampling. First, we use Monte Carlo simulations of models based on measurements to show that the methods are reliable and offer substantial improvements over previous methods. Next, we apply the methods on measurements of transcription intervals of different promoters in live *E. coli*, and show that they produce significantly different results, both in low and high noise settings, and that, in the latter case, they even lead to qualitatively different results. Finally, we demonstrate that the methods can be generalized for other similar purposes, such as for estimating gene activation kinetics. In this case, the new methods allow quantifying the inducer uptake dynamics as opposed to just comparing them between cases, which was not previously possible. We expect this new methodology to be a valuable tool for functional analysis of cellular processes using single-molecule or single-event microscopy measurements in live cells.

Availability: Source code is available under Mozilla Public License at http://www.cs.tut.fi/%7Ehakkin22/censored/.

Contact: andre.ribeiro@tut.fi

1 INTRODUCTION

In bacteria, transcription is the main regulatory mechanism of RNA numbers in the cell. This conclusion is supported by the lack of correlation between RNA numbers and their degradation rates (Bernstein *et al.*, 2002) and by the fact that most regulatory molecules modulate the transcription initiation process (McClure, 1985). Relevantly, the regulatory mechanisms of transcription allow wide adaptability, as the kinetics of this process varies widely between promoters, and for the same promoter under different conditions.

Live cell measurements in *Escherichia coli* suggest that, depending on the promoter and the conditions, such as the presence/absence of repressor/activator molecules, RNA production can range from sub-Poissonian, that is, less uncertain than a Poisson process (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012), through Poissonian (Yu *et al.*, 2006), to super-Poissonian (Golding *et al.*, 2005; Taniguchi *et al.*, 2010). Such wide dynamic range is not likely achievable by a single mechanism and, in agreement, evidence suggests that this a multi-step process (McClure, 1985).

Currently, the subprocesses that constitute transcription cannot be directly measured in live cells. However, it is possible to observe RNA production of individual promoters with single molecule resolution over time (Golding *et al.*, 2005). This information can be used to estimate the dynamical parameters of the subprocesses by the means of a stochastic model (Kandhavelu *et al.*, 2011). The success of this strategy requires accurate and unbiased statistical methods of data analysis as well as a model that can account for all possible dynamical regimes.

Previously, we made use of distributions of intervals between transcription events in various conditions in order to estimate in maximum likelihood sense, for each condition, the number and duration of the rate limiting steps in transcription (Kandhavelu *et al.*, 2011). Relevantly, unlike RNA numbers, these intervals are not affected by RNA degradation, or dilution due to cell division, and consequently allow more accurate quantification of the transcription process. However, the previous model of transcription does not cover all potential cases (e.g. super-Poissonian RNA production).

Here, we first propose a model that, by combining previous models responsible for different dynamical behaviors (McClure, 1985; Peccoud and Ycart, 1995), is capable of exhibiting behaviors ranging from sub- to super-Poissonian. Next, we present methods to estimate its parameters in maximum likelihood sense. An advantage over previous methods (Kandhavelu *et al.*, 2011) is that the new methods also use information of the unobserved transcription events. Such additional information results in improved accuracy and can be used to correct the biases resulting from the limited the measurement time. The methods can also account for the discrete sampling, which is typical for a fluorescence microscopy measurement. The increased accuracy allows studying subprocesses with smaller time scales, while the lack of bias is essential in order to correctly estimate the parameters of the more noisy models and to compare them in an unbiased manner. The methods can also be used to provide features such as confidence in the estimated parameters, and we use statistical methods to select components of the model in/out, which can be used to determine if certain components are responsible for the observed dynamical behavior.

*to whom correspondence should be addressed

2 METHODS

Transcription intervals and activation time measurements in live *E. coli* cells were obtained using the MS2-GFP RNA-tagging system (Golding *et al.*, 2005). The cells contain a single-copy vector coding for the target RNA under the control of a specific target promoter: TetA (Muthukrishnan *et al.*, 2012), bacteriophage λ RM (Golding *et al.*, 2005), or arabinose BAD (Makela *et al.*, 2013). The target RNAs contain an array of 48 or 96 binding sites (depending on the construct) for the highly expressed MS2-GFP reporters to bind. This allows the target RNAs to be visualized using fluorescence microscopy right after their production. The systems were constructed previously (Golding *et al.*, 2005; Muthukrishnan *et al.*, 2012; Makela *et al.*, 2013), and more details of the measurements conducted here using these systems are given in the supplement.

3 ALGORITHMS

In this section, we propose a model of transcription initiation whose kinetics can, depending on parameters selection, range from sub- to super-Poissonian. Next, we describe how to extract time interval distributions from the model. Finally, we describe how to estimate the model parameters using the measurement data.

3.1 Model of transcription initiation

To allow transcription dynamics to range from sub- to super-Poissonian, we propose the model of elementary reactions:

$$\underbrace{P_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} P_{\text{on}} \xrightarrow{k_1} I_1}_{\text{on-off step}} \xrightarrow{k_2} \cdots \xrightarrow{k_{n-1}} I_{n-1} \xrightarrow{k_n} P_{\text{on}} + E \qquad (1)$$

where P_{off}, P_{on}, $I_j \in \mathbb{Z}_1$ represent different states of the promoter: inactive, active, and some intermediate states of transcription initiation, respectively, while E represents the elongation complex. This model is designed to combine the active-inactive promoter model (Peccoud and Ycart, 1995) with a sequential model of transcription initiation (McClure, 1985; Saecker *et al.*, 2011). Note that any number of backward reactions for steps 1 through $n-1$ is implicitly supported, since equal dynamics can be achieved by setting the rates of the model appropriately (see supplement).

The on-off mechanism produces bursty RNA production, due to the random off periods (Peccoud and Ycart, 1995). When this mechanism dominates the dynamics, the intervals between transcript production are highly noisy, resulting in super-Poissonian RNA production. The above model is appropriate regardless of the mechanism controlling the promoter on-off transitions as long as the state transitions occur with constant probability per unit time. Recent studies suggest that the dynamics of several promoters in *E. coli* may be dominated by such a mechanism (Golding *et al.*, 2005; Taniguchi *et al.*, 2010; Chong *et al.*, 2014).

In contrast, a sequential process of RNA production reduces noise, as it produces more regular intervals. *In vitro* measurements suggest the following sequence of events (McClure, 1985; Lutz *et al.*, 2001): first, an RNA polymerase must find and diffuse along the DNA template until finding the transcription start site (Saecker *et al.*, 2011). There, the polymerase forms a closed complex, and then goes through several isomerization steps, until completing the open complex formation (Saecker *et al.*, 2011). After escaping the start site, the complex elongates along the DNA template, clearing the promoter region. Recent *in vivo* measurements have shown that at least some promoters in *E. coli* are capable of exhibiting a dynamics consistent with this model for a wide range of conditions (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012).

In Equation (1), the steps which are much faster than the others can be neglected. Using the same argument, we can also take elongation complexes E to represent fully transcribed RNAs, as elongation takes tens of seconds (Herbert *et al.*, 2006), while inter-production intervals are in the order of hundreds of seconds (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012). Regardless, elongation is not expected to affect the RNA production intervals on average, unless the initiation rate is so high that there is polymerase traffic (Rajala *et al.*, 2010). We also note that if the promoter states are unobservable (as is the case here), the order of sequential processes cannot be determined from the transcription dynamics.

3.2 Transcription interval distribution

Since transcription intervals can be split to sums of independent steps, the probability densities are easily manipulated in terms of their moment generating functions (MGFs). This is due to the fact that the MGF of a sum of independent variables is the product of the individual MGFs, and the MGF of a mixture is a weighted sum of the individual MGFs. The MGF of an exponential variate with a mean of ${k_i}^{-1}$ is:

$$M_i(t) = k_i \ (k_i - t)^{-1} \qquad (2)$$

which implies that the MGF of the on-off step (including on, off, and the first reaction) is:

$$\begin{aligned} &M_{\text{on-off}}(t) = k_1 \ (k_{\text{on}} - t) \ \left(p^- - t\right)^{-1} \left(p^+ - t\right)^{-1} \\ &\text{with} \quad p^{\pm} = \tfrac{k_{\text{off}} + k_1 + k_{\text{on}}}{2} \pm \tfrac{\sqrt{(k_{\text{off}} + k_1 - k_{\text{on}})^2 + 4\, k_{\text{off}}\, k_{\text{on}}}}{2} \end{aligned} \qquad (3)$$

which is described in more detail in the supplement. This indicates that any MGF of the model must have the form:

$$M(t) = G \prod_{i=1}^{U} (z_i - t)^{u_i} \prod_{i=1}^{V} (p_i - t)^{-v_i} = \sum_{i=1}^{V} \sum_{v=1}^{v_i} \frac{R_i^{(v)}}{(p_i - t)^v} \qquad (4)$$

where the latter is the partial-fraction decomposition of $M(t)$. The residues $R_i^{(v)}$ can be computed e.g. using: $z - t = (z - p) + (p - t)$ and $1/(p-t)/(q-t) = -1/(p-q)/(p-t) + 1/(p-q)/(q-t)$. By noting that the decomposition specifies a linear combination of the MGFs of sums of exponential variates, the probability density (PDF) is recovered as:

$$f(x) = \sum_{i=1}^{V} \sum_{v=1}^{v_i} \left(\frac{R_i^{(v)}}{{p_i}^v} \right) \frac{{p_i}^v}{\Gamma(v)} \, x^{v-1} \, e^{-p_i \, x} \qquad (5)$$

where the parenthesized term is the mixing weight and the remaining term is the PDF of a sum of v exponential variates, with a mean of ${p_i}^{-1}$ each. Here, $\Gamma(w)$ denotes the factorial of $w-1$. In general, the mixing weights are not convex, so this is not a proper mixture density.

Manipulating the MGF can be also used to perform other useful operations: the survival function can be obtained by adding a pole at $t = 0$, subtracting the t^{-1} term, and taking the inverse transform as above. Also, differentiation might be easier to perform on the MGF, which is useful for evaluating gradients and/or Hessians for optimization or for confidence estimation. For example, differentiation with respect to a k_i^{-1} with $i > 1$ can be achieved by adding a zero at $t = 0$, a pole at $t = k_i$, and multiplying the residues by k_i.

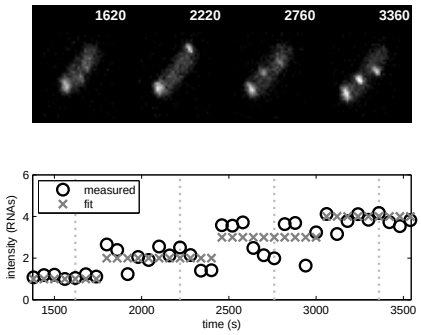

Fig. 1. Example data from MS2-GFP-tagged RNA measurements with the tetA promoter. Upper panel: fluorescence microscopy images of a cell at different time points, as indicated by the time stamp (seconds). Lower panel: extracted intensities and estimated RNA numbers of the cell shown in the upper panel. The vertical lines represent the time points in the upper panel.

3.3 Maximum likelihood estimation

As the measurements have finite length and discrete sampling (see Figure 1), the true intervals between RNA production are not exactly known. An interval can be only observed if it fits in to the measurement window, and as such, in each cell, the last interval will not be observed due to the end of the measurement period. Neglecting these unobserved intervals will result in underestimation of the interval durations. Meanwhile, the discrete sampling implies that each interval contains some uncertainty about its exact duration, which should be communicated to the estimator. For example, the true interval between the second and third production in Figure 1 is known to be 10 to 12 units long (interval-censoring). Meanwhile, the true interval between the third and fourth production is no less than 9 units long (right-censoring). These two modes of uncertainty are called interval- and right-censoring, as the true value is bounded to an interval or to the right (on the real line) of some observation. More precise definitions can be found e.g. in Turnbull (1976).

Provided that the intervals T_i between transcription events are independent, the probability of observing a sequence of intervals $(t_1, \cdots, t_m)$ in a time series of length L is given by:

$$\begin{aligned} &\mathbb{P}[\, S \simeq (t_1, \cdots, t_m)\,] = \mathbb{P}[\, T_1 \simeq t_1\,] \cdots \mathbb{P}[\, T_m \simeq t_m\,] \\ &\quad \mathbb{P}[\, T_{m+1} > L - (t_1 + \cdots + t_m)\,]\; \mathbb{I}[\, t_1 + \cdots + t_m \leq L\,] \end{aligned} \tag{6}$$

where the notation abuses $\mathbb{P}[\, X \simeq x\,] = \mathbb{P}[\, x \leq X < x + \partial x\,]$ for infinitesimal ∂x and $\mathbb{I}[\,\cdot\,]$ is the indicator function. Here, m and L need not to be constant, but can be realizations of independent random variables. This implies that a maximum likelihood (ML) estimator can be written in the following form:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}}\; \mathbb{P}[\, t_1 \in [x_1, y_1], \cdots, t_{m+1} \in [x_m, y_m]\,|\,\boldsymbol{\theta}\,] \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m+1} \log(\, F(y_i\,|\,\boldsymbol{\theta}) - F(x_i\,|\,\boldsymbol{\theta})\,) \end{aligned} \tag{7}$$

where x_i and y_i are the (possibly infinite) known bounds for t_i, and $F(x)$ is the cumulative density (CDF). The ordinary ML estimator is recovered at the limit $y_i \to x_i$, since $F(y_i\,|\,\boldsymbol{\theta}) - F(x_i\,|\boldsymbol{\theta}) \to f(x_i\,|\,\boldsymbol{\theta})\,(y_i - x_i)$, where $f(x)$ is the PDF, and $(y_i - x_i)$ is constant with respect to $\boldsymbol{\theta}$.

In general, the times from the beginning of the measurement to the first production might not have the appropriate distribution, and they cannot be used in the estimator. An exception to this occurs when it is known that the transcription process starts at the same time as the measurement. Another exception occurs when the transcription intervals are exponential, in which case the first production has the appropriate distribution due to the memorylessness of the exponential distribution.

If interval-censoring of the consecutive intervals is used, the samples are not independent, since the error terms of the consecutive measurements might be correlated. However, if the production intervals are much longer than the sampling intervals, which is necessary for accurate estimation anyway, these correlations tend to be negligible. Despite violating this assumption, we found interval-censoring to improve the estimator performance considerably, as shown below.

In some cases, simple solutions to the ML problem exist (see the supplement). However, the general ML problem requires numerical methods. Also, the ML surface is not guaranteed to be concave, nor even unimodal. However, it tends to be well-behaved in practice, especially for larger samples (the usual properties of an ML estimator apply). Due to this, we perform 100 restarts with random starting point. We used the Nelder and Mead (1965) method for optimization, since it appeared to perform well and is fast. We also experimented with the Broyden-Fletcher-Goldfarb-Shanno method, with either exact or finite difference derivatives, but it produced similar results and was significantly slower.

4 RESULTS

4.1 Applying the methods on Monte Carlo simulations

Throughout this manuscript, we use "exp" and "seq-n" to denote models of 1 or n sequential exponential steps and no on-off mechanism, respectively, and "onoff", and "onoff-n" to denote the full models with 1 or n steps, respectively. The parameters of each model is given as a vector $({k_{\text{on}}}^{-1}, {k_{\text{off}}}^{-1}, {k_1}^{-1}, \cdots, {k_n}^{-1})$ with ${k_2}^{-1} \leq \cdots \leq {k_n}^{-1}$, since the order of k_i is exchangeable.

We performed Monte Carlo simulations using the following models based previous live *E. coli* measurements: exp with a mean of 2750 s with 60 cells sampled every 180 s for 3300 s (Yu *et al.*, 2006), seq-2 with means of 712 and 716 s with 40 cells sampled every 60 s for 7200 s (Kandhavelu *et al.*, 2011), seq-3 with means of 109, 254, and 254 s with 113 cells sampled every 60 s for 3600 s (Muthukrishnan *et al.*, 2012), and onoff with ${k_{\text{on}}}^{-1} = 360$ s, ${k_{\text{off}}}^{-1} = 1020$ s, and ${k_1}^{-1} = 102$ s with 100 cells sampled every 20 s for 4800 s (Chong *et al.*, 2014). We constructed a hypothetical onoff-2 model based on the onoff model, by setting ${k_1}^{-1} = {k_2}^{-1} = 51$ s, since there are no live *E. coli* measurements supporting the more complex on-off models. However, such models have been used in eukaryotic context (Blake *et al.*, 2003).

The shapes of the model distributions are shown in Figure 2, and model statistics are shown in Table S1. The table shows the parameters, the mean and standard deviation (sd) of the transcription intervals, their noise, as determined by the squared coefficient of variation, and the differential entropy, which is useful in interpreting

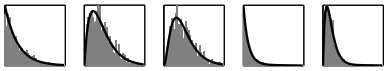

Fig. 2. Probability densities for the interval distributions of the models. The black curves show the asymptotic distribution, and the gray bars the histogram of 1000 random samples. From left to right, the models are: exp, seq-2, seq-3, onoff, and onoff-2.

the entropy-based statistics. In addition, the number of cells and the sampling (samples $\times$ sampling interval) is shown, which apply for the time series simulations.

4.1.1 Model selection and effects of sample sizes. First, we generated 100, 500, or 1000 intervals from each model, and fit the data with the following models: exp, seq-2, seq-3, onoff, onoff-2, onoff-3. A network of likelihood ratio (LR) tests is used to choose the preferred model, such that the least complex model which cannot be rejected at a significance level of 0.01 is selected. Table S2 shows the statistics for the most frequently selected model. Statistics were gathered from 1000 simulations in each case.

In the tables, log-likelihood (LL) quantifies how well the estimated model fits the data. The differential Kullback-Leibler divergence (KLD) from the true model to the estimated one measures the information lost when the estimated model is used to approximate the true model, which will, as opposed to the likelihood, penalize overfitting. The spatial median (med) of the parameter estimates represents a typical estimate. Finally, the choice frequency indicates how often this model is selected in favor of the others. The alternative model (alt model) indicates the second most frequently selected model.

The results indicate that the information lost with the on-off models is an order of magnitude greater, which is expected, since they are more noisy. The likelihood values suggest that the estimators behave as is expected from an ML estimator, despite the numerical optimization procedure. For the exp and seq-2 models, 100 samples appears to be sufficient to identify the appropriate model for more than 90% of the time. For 500 or 1000 samples, this is true for all but the onoff model, in which it occurs more than 80% of the time. Finally, we note that especially with 100 samples, the multi-parameter models have biases in the parameter estimates, such as towards/away from equal values in the sequential models. Fortunately, these biases vanish with larger sample sizes, as is typical for maximum likelihood estimates, so the problem can be mitigated by collecting more samples.

4.1.2 Advantages of censoring. Next, we generated time series of RNA numbers to simulate our measurement settings. The data extracted from these series were fit with the appropriate model using full censoring, without interval-censoring (i.e. disregarding discrete sampling effects), without right-censoring (i.e. without correcting for the unobserved samples), or with neither mode of censoring. For the sequential exponential models, the last method corresponds to a previous method (Kandhavelu *et al.*, 2011).

The results are shown in Table S3. Again, the log-likelihood quantifies how well the estimated model fits the data, the divergence how well the estimated model corresponds to the true model, and the median of the estimates represents a typical parameter estimate. The likelihood values are not comparable between the different modes of censoring. While these results demonstrate the applicability of the methods under typical settings, the different models cannot be compared as the relative sampling settings of the models differ widely (cf. Samples in Table S3). To allow such comparison, we also simulated each model for $5\,\mu$ time units, sampling every $5^{-1}\,\mu$ units, where μ is the mean production interval. These results are shown in Table 1.

The results indicate that the lack of right-censoring will result in drastic underestimation of the transcription interval duration, especially in the noisy cases, such as with the on-off models. Also, without right-censoring, the variance is generally underestimated in a similar manner. In the tested settings, the effects of discrete sampling were found comparatively weaker, but this is likely mitigated by the relatively frequent sampling (sampling rates are around 10-fold to that of transcription). However, accounting for it offers slight improvements in the accuracy in all cases.

In summary, it is essential to apply both modes of censoring for the high-noise models, while for sequences of exponentials neither mode of censoring is critical. However, in all cases the variant with full censoring performs best, as expected. Also, if the true model is not known but model selection will be performed, it is again necessary to use censoring to avoid biases towards selecting a less noisy model.

4.2 Applying the methods on transcription interval measurements

4.2.1 Transcription kinetics of the tetA promoter. We used MS2-GFP RNA-tagging and the methods with full and no censoring to analyze transcript production intervals in live *E. coli* (see Figure 1). This allows determining if the two methods result in significantly different results with true measurement data. For this, we performed experiments, as described in the Methods section. Again, the model is selected using a network of LR tests and the following models: exp, seq-2, seq-3, onoff, onoff-2, onoff-3.

First, we measured the RNA production in a construct where the target gene is controlled by the tetA promoter. A previous study reports that the dynamics are explained by a sequence of two or three exponentials, depending on the conditions (Muthukrishnan *et al.*, 2012). Our measurements were conducted under full induction (15 ng/ml of anhydrotetracycline) (Muthukrishnan *et al.*, 2012), imaging the cells for every 1 min for a duration of 60 min.

The histogram of observed intervals collected from the measurements is shown in the upper panel of Figure S1 along with the PDFs of the estimated models. The intervals were extracted by analyzing the time series of individual cells separately (as opposed to observing production intervals in the whole cell lineage). By pooling the observed intervals from multiple cells, we implicitly assume that there are no significant variations in the model parameters between the cells. In addition, the lower panel of Figure S1 shows the Turnbull's CDF estimate (Turnbull, 1976), which is a nonparametric maximum likelihood estimate of the CDF accounting for both modes of censoring. Note that the histogram only contains the observed samples, so it is expected to underestimate (overestimate) the probability for large (small) values. On the other hand, the Turnbull estimator is expected to well represent the true CDF.

Table 2 shows statistics for the two methods. The number of samples is different in the two cases, as the full censoring method also

Table 1. Performance when using different modes of censoring in Monte Carlo simulations.

Model	exp	seq-2	seq-3	onoff	onoff-2
Samples mean (sd)	500 (22.1)	474 (16)	468 (13.7)	562 (29.1)	514 (20.1)
KLD mean (sd)	0.00161 (0.00214)	0.00184 (0.00227)	0.00273 (0.00242)	0.0247 (0.0762)	0.042 (0.0466)
LL mean (sd)	−792 (28.5)	−698 (20.1)	−647 (17.7)	−859 (35.2)	−750 (24.2)
Parameter med	2680	694, 714	162, 190, 257	319, 812, 96.7	464, 926, 51.7, 48.6
Samples mean (sd)	500 (22.1)	474 (16)	468 (13.7)	562 (29.1)	514 (20.1)
KLD mean (sd)	0.0019 (0.00252)	0.00282 (0.00321)	0.00385 (0.00356)	0.0325 (0.0485)	0.0455 (0.0483)
LL mean (sd)	−3590 (180)	−3070 (118)	−2680 (89.4)	−2700 (145)	−2360 (96.9)
Parameter med	2860	644, 820	105, 247, 278	566, 1120, 107	828, 999, 55.6, 48.1
Samples mean (sd)	401 (22)	374 (16)	368 (13.7)	465 (28.5)	416 (19.7)
KLD mean (sd)	0.0378 (0.0132)	0.0209 (0.0101)	0.0178 (0.00915)	0.178 (0.0669)	0.142 (0.0408)
LL mean (sd)	−3470 (179)	−2990 (118)	−2610 (90.4)	−2540 (144)	−2250 (98.7)
Parameter med	2120	620, 623	119, 214, 223	0.415, 112, 81.2	41.9, 559, 45.1, 44.3

Blocks from top to bottom: full censoring, no interval-censoring, and no censoring. The table shows the mean (sd) number of samples per time series, the mean (sd) Kullback-Leibler divergence (KLD), the mean (sd) log-likelihood (LL), and the spatial median of the parameter estimates. Units of time are in seconds, and the entropy-based measures are in nats.

Table 2. Statistics of the estimated models for the tetA promoter.

Method	Full censoring	No censoring
Samples	362	254
Sel model	seq-3	seq-3
Parameters	131, 131, 522	109, 254, 254
Parameter sd	109, 109, 54.2	119, 138, 139
Est mean (sd)	784 (554)	617 (375)
Est cv-squared	0.499	0.370
Samples	345	175
Sel model	seq-3	seq-3
Parameters	131, 131, 522	171, 171, 171
Parameter sd	171, 173, 76	134, 136, 137
Est mean (sd)	784 (554)	514 (297)
Est cv-squared	0.499	0.333

Blocks from top to bottom: 60 min series and 30 min series. The table shows the number of samples, the selected model (sel model), the estimated parameters and estimates of their standard deviation, and the mean (est mean), standard deviation (est sd) and squared coefficient of variation (est cv-squared) resulting from the estimated model.

includes the right-censored samples. Here, both methods suggest a three-exponential model and rule out the possibility of an on-off mechanism as the primary regulator of the dynamics. In both cases, the estimated parameter standard deviation is about 2 min per parameter, with full censoring resulting in a slightly higher confidence on the parameters. With full censoring, there is about 1.3-fold increase in the mean, which is expected, since neglecting the right-censored data tends to result in underestimation of the durations. To confirm the statistical significance of this difference, we performed a one-sided t-test with a null hypothesis that the two means are equal, resulting in a p-value of 5.20×10^{-9}. Also, we found differences in the noise levels, but both methods suggest cv-squared of less than 0.5, favoring the three-exponential model.

To demonstrate that the full censoring method is immune to changes in the measurement duration, we repeated the estimation procedures such that the time series was split into two halves of 30 min each, from which the data were extracted separately. The results are shown in the lower block of Table 2, and they indicate that with the full censoring method, only the confidence in the parameter estimates is visibly affected, whereas the non-censoring method underestimates the mean and standard deviation (even more than when applied to the time series 60 min long).

Finally, we studied whether our results are affected by the elongation process. First, to test if significant RNA polymerase traffic occurs (e.g. due to pausing), we estimated the correlation between consecutive transcription intervals. We found the correlation to be 0.149 with a p-value of 0.196 in a LR test with a null model of uncorrelated normal data, indicating that the correlation is not significant. Next, we added a normal zero-mean noise term to the transcription model to simulate stochasticity resulting from chain elongation. The estimated sd of this noise term was 24.4 s (p-value of 0.574 in an LR test with a null model from Table 2), suggesting that such noise term is not significant at our resolution. As neither of the null hypotheses can be rejected, we conclude that there is no evidence that the dynamics of elongation affects our results. In addition, in agreement, we note that no differences have been found in the dynamics of RNA production between constructs with 48 and 96 of the MS2-GFP binding sites (Golding *et al.*, 2005; Hakkinen *et al.*, 2014).

4.2.2 Transcription kinetics of the λ RM promoter. Next, we analyzed measurements of the transcription intervals of the MS2-GFP-tagged target RNA controlled by the bacteriophage λ RM promoter. This construct is expected to result in a bursty, highly noisy expression (Golding *et al.*, 2005). In this case, the cells were imaged every 1 min for a duration of 120 min.

Again, the histogram of the observed intervals and the PDFs of the estimated models are shown in the upper panel of Figure S2, and the Turnbull's CDF estimate and the CDFs of the models are

Table 3. Statistics of the estimated models for the bacteriophage λ RM promoter.

Method	Full censoring	No censoring
Samples	303	155
Sel model	onoff	onoff-2
Parameters	5840, 1730, 1140	782, 2450, 527, 25.5
Parameter sd	3020, 469, 134	532, 2880, 90.7, 11.0
Est mean (sd)	4990 (8360)	721 (864)
Est cv-squared	2.81	1.44
Est burst size (interval)	1.52 (7560)	4.65 (3350)
Est duty cycle	0.228	0.768

The table shows the number of samples, the selected model (sel model), the estimated parameters and estimates of their standard deviation, and the mean (est mean), standard deviation (est sd) and squared coefficient of variation (est cv-squared) resulting from the estimated model. Also shown is the burst size, burst interval, and the duty cycle of the estimated model.

shown in the lower panel. Similarly, Table 3 shows statistics from the two estimation procedures.

Both methods indicate that an on-off model is required to explain the measurements, and that the interval distribution is highly noisy (cv-squared above unity). However, in the case where the unobserved samples (which constitute around 50% of the samples) are neglected, both the mean and sd of the distribution are drastically (over 5-fold) underestimated. In terms of cv-squared, this results in a 2-fold underestimation of stochasticity in the transcription initiation process.

Finally, we computed the statistics of the bursts in the two estimated models, which are shown in Table 3. The model estimated using full censoring suggests that the noise results from a low duty cycle (i.e. the gene being repressed most of the time), while the model estimated without censoring suggests that the noise is due to the large size of the bursts.

4.3 Applying the methods on measurements of external transcription activation times

Finally, we analyzed the activation dynamics of the arabinose BAD promoter to demonstrate that the methods generalize to other estimation problems. The analysis was performed by collecting both the time for each cell to produce the first RNA after introducing arabinose in the medium, and the subsequent time intervals between consecutive productions of transcripts. For this, 1% of L-arabinose was introduced at the start of the measurement, after which cells were imaged every 1 min for 120 min.

The time to produce the first RNA is expected to include the time for the cell to uptake sufficient arabinose to turn on the active arabinose uptake system, for the arabinose to form the complex activating the BAD promoter (Daruwalla *et al.*, 1981), and for the first transcript to be produced. As such, after Megerle *et al.* (2008) and Makela *et al.* (2013), we fit the following model:

$$\underbrace{A_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftarrows}} A_{\text{on}} \xrightarrow{k_1} P_{\text{on}}}_{\text{activation}} \underbrace{\xrightarrow{k_2} I_2 \xrightarrow{k_3} I_3 \xrightarrow{k_4} P_{\text{on}} + E}_{\text{transcription}} \tag{8}$$

Table 4. Statistics of the estimated models for the BAD promoter.

Distribution	First production	Activation	Transcription
Samples	599	-	345
Model	seq-5	seq-2	seq-3
Parameters	(see right)	33.2, 1580	9.64, 1110, 1750
Parameter sd	(see right)	124, 201	39.1, 469, 512
Est mean (sd)	4490 (2610)	1620 (1580)	2870 (2080)
Est cv-squared	0.338	0.960	0.521

The table shows the number of samples, the selected model (sel model), the estimated parameters and estimates of their standard deviation, and the mean (est mean), standard deviation (est sd) and squared coefficient of variation (est cv-squared) resulting from the estimated model.

where A_{on}, $A_{\text{off}} \in \mathbb{Z}_1$ represent the states that the internal arabinose concentration has or has not reached sufficient concentration to turn on the arabinose uptake mechanism, respectively, and P_{on}, $I_j \in \mathbb{Z}_1$ represent the states of the BAD promoter. Since the intervals were found to have low noise (cv-squared of 0.347), we model transcription with a sequence of three exponentials.

In the above model, the times for the first RNA production follow an seq-5 model with the parameters (p^-, p^+, k_2, k_3, k_4), where $p^{\pm}$ are as in Equation (3), and the intervals of the subsequent productions follow a seq-3 model with the parameters (k_2, k_3, k_4). As the models share parameters, they are fit jointly to both data. The histogram of the observed data and the PDFs of the estimated models are shown in the upper panels of Figure S4, and the Turnbull's CDF estimates and the CDFs of the models are shown in the lower panels. Table 4 shows statistics from the estimation procedure. The exponential-likeness of the activation process suggests that either k_{on}, k_1, or both must be fast (non-rate limiting). Meanwhile, in transcription, two of the rates k_2, k_3, and k_4 are rate limiting.

The mean (sd) of the observed first production times and transcription intervals were 3880 s (1700 s) and 1700 s (1000 s), respectively, which agrees with those reported in (Makela *et al.*, 2013). Again, this suggest that neglecting the unobserved data results in slight underestimation of the mean and the variance. Regardless, the qualitative results, such as noise, reported in (Makela *et al.*, 2013) appear to hold.

It is worth noting that, to avoid artificial correlations between the first and the subsequent production times, Makela *et al.* (2013) used a windowing method to compare the activation dynamics in different conditions. In our method, such windowing is not needed, as the censored data is used. Because of this, in addition to using more information, our method also allows unbiased quantification of the different distributions, not just their comparison. Furthermore, our method can also be used to deconvolve the activation dynamics distribution, as shown in Table 4.

5 DISCUSSION

We have proposed a model that combines a promoter on-off mechanism (Peccoud and Ycart, 1995) with a sequential process of transcription initiation (McClure, 1985), which allows explaining recent measurements of transcription dynamics under a wide range

of conditions (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012; Yu *et al.*, 2006; Golding *et al.*, 2005; Taniguchi *et al.*, 2010), and established methods to estimate its parameters in maximum likelihood sense using transcription interval data.

The methods enable more accurate quantification of the transcriptional dynamics both in theory and in practice, as demonstrated by the Monte Carlo simulations, as well as testing if particular components of the model are responsible for the observed dynamics. In addition, we compared the methods with previous methods using measurement data from live *E. coli*, and showed that the new methods produce significantly different results and can provide new biological insight (e.g. on the underlying sources of noise in transcription). Finally, we demonstrated that the methods have a wider applicability on problems of similar nature, such as estimating the kinetics of external activation of a promoter.

In its present form, the proposed model should already be detailed enough to allow a genome-wide analysis of transcription and transcription activation of individual genes under a wide range of conditions in prokaryotes, which is necessary to understand how inducers and repressors regulate the dynamics of gene expression. Further, we believe that our methods can be extended to enable future studies of eukaryotic transcription dynamics and of translational dynamics at the single protein level.

ACKNOWLEDGEMENTS

Funding: Work supported by Jenny and Antti Wihuri Foundation [to A.H.] and Academy of Finland [257603 to A.S.R.].

REFERENCES

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N. (2002). Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **99**(15), 9697–9702.

Blake, W. J., Kaern, M., Cantor, C. R., and J., C. J. (2003). Noise in eukaryotic gene expression. *Nature*, **422**(6932), 633–637.

Chong, S., Chen, C., Ge, H., and Xie, X. S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell*, **158**(2), 314–326.

Daruwalla, K. R., Paxton, A. T., and Henderson, P. J. F. (1981). Energization of the transport systems for arabinose and comparison with galactose transport in *Escherichia coli*. *Biochem. J.*, **200**(3), 611–627.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**(6), 1025–1036.

Hakkinen, A., Kandhavelu, M., Garasto, S., and Ribeiro, A. S. (2014). Estimation of fluorescence-tagged RNA numbers from spot intensities. *Bioinformatics*, **30**(8), 1146–1153.

Herbert, K. M., La Porta, A., Wong, B. J., Mooney, R. A., Neuman, Keir C. adn Landick, R., and Block, S. M. (2006). Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*, **125**(6), 1083–1094.

Kandhavelu, M., Mannerstrom, H., Gupta, A., Hakkinen, A., Lloyd-Price, J., Yli-Harja, O., and Ribeiro, A. S. (2011). In vivo kinetics of transcription initiation of the lar promoter in *Escherichia coli*. evidence for a sequential mechanism with two rate-limiting steps. *BMC Syst. Biol.*, **5**, 149.

Lutz, R., Lozinski, T., Ellinger, T., and Bujard, H. (2001). Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator. *Nucl. Acids Res.*, **29**(18), 3873–3881.

Makela, J., Kandhavelu, M., Oliveira, S. M. D., Chandraseelan, J. G., Jason, L.-P., Peltonen, J., Yli-Harja, O., and Ribeiro, A. S. (2013). *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucl. Acids Res.*, **41**(13), 6544–6552.

McClure, W. R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, **54**, 171–204.

Megerle, J. A., Fritz, G., Gerland, U., Jung, K., and Radler, J. O. (2008). timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys. J.*, **95**(4), 2103–2115.

Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., Kudasov, F., Chowdhury, S., Yli-Harja, O., and Ribeiro, A. S. (2012). Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucl. Acids Res.*, **40**(17), 8472–8483.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Comput. J.*, **7**(4), 308–313.

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.*, **48**(2), 222–234.

Rajala, T., Hakkinen, A., Healy, S., Yli-Harja, O., and Ribeiro, A. S. (2010). Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput. Biol.*, **6**(3), e.1000704.

Saecker, R. M., Record, Jr., M. T., and deHaseth, P. L. (2011). Mechanism of bacterial transcription initiation. *J. Mol. Biol.*, **412**(5), 754–771.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**(5991), 533–538.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Royal Stat. Soc. Ser. B*, **38**(3), 290–295.

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X. S. (2006). Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**(5767), 1600–1603.

S1 SUPPLEMENTARY INFORMATION

S1.1 MS2-GFP-tagged RNA measurements

S1.1.1 Cells, plasmids, and growth conditions for the tetA construct. The *E. coli* strain DH5α-PRO cells containing the MS2-GFP RNA-tagging system was engineered by Golding and Cox (2004) (generously provided by I. Golding, University of Illinois, USA), and the construct with the tetA promoter by Muthukrishnan *et al.* (2012). The cells contain a single-copy BAC vector coding for the mRFP1-MS2-96bs target RNA (Golding and Cox, 2004), controlled by the tetA promoter, as well as a low-copy pZS12MS2-GFP plasmid carrying the MS2-GFP reporter, controlled by the LlacO1 promoter (Le *et al.*, 2005) (generously provided by P. Cluzel, University of Chicago, USA). In the cells, the *tetR* gene, controlled by the N25 promoter, is integrated into the chromosome, allowing full range of induction of the target gene, while the *tetA* gene is absent. When a target RNA is produced in the cells, the abundant MS2-GFPs rapidly occupy the binding sites, allowing the visualization of the complex using fluorescence microscopy (see the upper panel of Fig. 1).

For overnight cultures, the strain from glycerol stock was inoculated in lysogeny broth (LB), with 10 g/l of tryptone (Sigma-Aldrich, USA), 5 g/l of yeast extract (LabM, UK), and 5 g/l of NaCl (LabM, UK), pH 7.0, with appropriate antibiotics (100 µg/ml ampicillin and 35 µg/ml chloramphenicol; both from Sigma-Aldrich, USA), and incubated at 37 °C with shaking (250 rpm). From the overnight cultures, the cells were inoculated into a fresh LB medium supplemented with antibiotics, with initial optical density (OD) of 0.1 at 600 nm, and incubated at 37 °C to mid-logarithmic phase with OD of 0.5. To induce the production of MS2-GFP proteins, 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG; Sigma-Aldrich, USA) was added to the medium at OD of 0.35, while the target was induced by adding anhydrotetracycline (aTc; Sigma-Aldrich, USA) to the liquid culture. After 5 min, the cells were placed on a microscope slide between a coverslip and 1% LB-agarose gel with IPTG (1 mM) and aTc (15 ng/ml) to maintain induction under the microscope. The image acquisition started around 20 min after induction of the target gene (including the 5 min in liquid culture). This interval suffices to reach a steady induction of the reporter (Le *et al.*, 2005).

S1.1.2 Cells, plasmid, and growth conditions for the λRM construct. The *E. coli* strain DHα-PRO cells with the λ P_{RM} reporter were generously provided by I. Golding (University of Illinois, USA) (Golding *et al.*, 2005). The cells contain the reporter P_{LtetO1}-MS2d-GFP and a single-copy target plasmid pIG-BAC (P_{RM} λ_{imm}(*rexAB*::bs48)), which contains an autoregulatory system coding for CI, controlled by the P_{RM} promoter, and Cro, controlled by the P_R promoter (Golding *et al.*, 2005). Further, the plasmid contains the immunity region of the wild-type λ with the *rexA* and *rexB* genes replaced with a 48 binding site array for MS2d proteins (Golding *et al.*, 2005), which allow tagging the RNAs produced by P_{RM} promoter. Depending on the occupation of the operator sites OR1, OR2, and OR3, either the P_{RM} or P_R promoter will be repressed (Svenningsen *et al.*, 2005), resulting in target RNA bursts (Golding *et al.*, 2005).

The cells were grown in LB medium with the following components: 10 g/l of tryptone, 5 g/l of yeast extract and 10 g/l of NaCl, with addition of 34 µg/ml of kanamycin and 34 µg/ml of chloramphenicol (both from Sigma Aldrich, USA). Initially, the cells were grown overnight with shaking at 260 rpm in an orbital shaker (Labnet) at 30 °C for 12 to 16 h to an OD of 0.1 at 600 nm. After this, they were grown to OD of around 0.01, diluted to 1:10 in LB medium with antibiotics, and further grown at 37 °C with shaking at 260 rpm for a few hours to reach exponential phase and OD of around 0.3.

The reporter gene was activated using 10 ng/ml of aTc, for at least 45 min, to allow the production and maturation of enough MS2-GFP proteins. For acclimatization, cells were grown at room temperature for 1 h. Next, 100 µl of melted agarose-medium with 1% agarose (Sigma life science, USA), LB medium, and 10 ng/ml aTc was poured onto a microscope slide with a glass coverslip on top. After the gel pad had solidified, the coverslip was removed and the gel pad was left to dry for 2 to 5 minutes at room temperature. Finally, 5 to 8 µl of cell suspension was added into the gel prior to imaging.

S1.1.3 Cells, plasmids, and growth conditions for the arabinose BAD promoter. The cells with the target gene controlled by the arabinose P_{BAD} promoter was engineered by Makela *et al.* (2013) (from the original construct generously provided by I. Golding, University of Illinois, USA). The *E. coli* strain DH5α-PRO cells contain the construct PROTET-K133, carrying $P_{LtetO-1}$ – -MS2d-GFP (Golding and Cox, 2004), along with the target construct, pMK-BAC (P_{BAD}-mRFP1-MS2-96bs), which is a single-copy F-based vector coding for a red fluorescent protein (mRFP1) followed by a 96 MS2d-GFP binding sites, controlled by P_{BAD} (Makela *et al.*, 2013). The DH5α-PRO strain is a native producer of AraC (Lutz and Bujard, 1997).

Cells were grown overnight at 30 °C with aeration and shaking in an LB medium, supplemented with antibiotics according to the plasmids. Next, cells were diluted into a fresh M63 medium and were allowed to grow to OD of 0.3 to 0.5 at 600 nm. To obtain full induction of the reporter, the cells were preincubated for 40 min with 100 ng/ml of aTc. Next, the cells were pelleted and resuspended in around 50 µl of fresh M63 medium, and few microliters of cells were placed between a 3% agarose gel pad made with medium and a glass coverslip. During imaging, a flow of fresh, pre-warmed M63 medium containing the inducer was provided using a peristaltic pump at a rate of 1 ml/min.

S1.1.4 Microscopy. Microscopy was performed using a Nikon Eclipse (TE2000-U; Nikon, Japan) inverted microscope with C1 confocal laser-scanning system and a 100× Apo TIRF (1.49 NA, oil) objective. The slides were kept in a temperature-controlled chamber (FCS2; Bioptechs, USA), which was pre-heated and kept at 37 °C. Prior to imaging, the cells were focused in a few seconds under light microscopy. The GFP fluorescence was excited using a 488 nm argon ion laser (Melles-Griot, USA) and measured using a 515/30 nm detection filter. Images were recorded every 1 min for 1 or 2 h using Nikon EZ-C1 software.

S1.1.5 Image processing. In all cases, the RNA production times are extracted from the microscopy images as follows (Kandhavelu *et al.*, 2012). First, a region occupied by each cell during the series is manually indicated. After this, the locations, dimensions, and orientations of the cells are found using principal component analysis, assuming that the fluorescence inside the cells is uniform. The RNA

spots are segmented using kernel density estimation with a Gaussian kernel. Next, the average intensity inside each cell but outside the RNA spots is computed to estimate the cell background intensity. The volume of the spots above this background is integrated to obtain the background-corrected total intensity for each cell at each time moment.

Since the lifetime of a tagged RNAs is much longer than the cell division time (Muthukrishnan *et al.*, 2012), this intensity is expected to be an increasing function, with a jump corresponding to an appearance of a tagged RNA. The jump positions are estimated by fitting a monotonic piecewise-constant curve in least-squares sense to the intensity series of each cell. The number of pieces is selected using an F-test with a p-value of 0.01, requiring higher-order curves to fit significantly better to justify their usage. See the lower panel of Fig. 1.

S1.2 Extended algorithms

S1.2.1 Transcription intervals of the on-off model. First, we consider the intervals of the state transition from S_1 to E (actually, a state transition from $(S_0, S_1, E) = (0, 1, E)$ to $(S_0, S_1, e) = (0, 0, e+1)$ in more proper terms) in the following model:

$$S_0 \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} S_1 \xrightarrow{k_2} E$$

Such transition takes a number of detours through state S_0 before transitioning to E through the reaction with rate k_2. The exit rate from S_1 is $v_1 \doteq k_{-1} + k_2$, and the probability of branching to S_0 from S_1 is $p_1 \doteq k_{-1} / v_1$. The MGF of the intervals of the state transition is the following linear combination:

$$M_1(t) = \left(\sum_{n=0}^{\infty} \underbrace{(1-p_1)\, {p_1}^n}_{\text{weight for } n \text{ detours}} \left(\underbrace{\frac{v_1}{v_1 - t}}_{M_{v_1}(t)} \underbrace{\frac{k_1}{k_1 - t}}_{M_{k_1}(t)} \right)^n \right) \times \underbrace{\frac{v_1}{v_1 - t}}_{M_{v_1}(t)}$$

where $M_{v_1}(t)$ and $M_{k_1}(t)$ are the MGFs for elementary transitions with rates v_1 and k_1, respectively. Provided that $v_1, k_1 > 0$, the geometric series converges in some open disk around $t = 0$, giving:

$$\begin{aligned} M_1(t) &= \frac{k_2\,(k_1 - t)}{(v_1 - t)\,(k_1 - t) - k_1\,k_{-1}} \\ &= \frac{k_1 - t}{k_1} \frac{p^-}{p^- - t} \frac{p^+}{p^+ - t} \end{aligned}$$

where the poles $p^{\pm}$ are the roots of $Q_1(t) \doteq k_1\,k_2 - (k_1 + k_{-1} + k_2)\,t + t^2$, and the coefficient is substituted using $Q_1(0) = k_1\,k_2 = p^-\,p^+$. From $M_1(t)$, it is apparent that the model is identifiable, that is, different choices of parameters k_1, k_{-1}, k_2 result in different interval distributions provided that the model is not degenerate.

Provided that $k_1, k_{-1} > 0$, the zeros and poles $p^- < k_1 < p^+$ are distinct, and the PDF is given by:

$$\begin{aligned} f_1(x) = &\underbrace{\frac{k_1 - p^-}{k_1} \frac{p^+}{p^+ - p^-}}_{\text{mixing weight}} \underbrace{p^-\,\exp(-p^-\,x)}_{\text{PDF}} + \\ &\frac{p^+ - k_1}{k_1} \frac{p^-}{p^+ - p^-}\, p^+\,\exp(-p^+\,x) \end{aligned}$$

which is a proper mixture of two exponential distributions, with rates p^-, p^+, where the former rate defines the tail behavior and the latter the behavior around $x = 0$.

Moreover, the ith cumulant is:

$$\kappa_i = \left[\frac{\partial^i}{(\partial t)^i} \log M_1(t) \right]_{t=0} = \Gamma(i) \left(-\frac{1}{{k_1}^i} + \frac{1}{(p^-)^i} + \frac{1}{(p^+)^i} \right)$$

from which the mean $\mu = \kappa_1$, variance $\sigma^2 = \kappa_2$, and higher-order moments can be obtained. By using the expansions of $Q_1(0)$ from above and $-Q_1'(0) = k_1 + k_{-1} + k_2 = p^- + p^+$, we have:

$$\mu = \left(\underbrace{\frac{k_1}{k_1 + k_{-1}}}_{\text{duty cycle}} \right)^{-1} \underbrace{\frac{1}{k_2}}_{\text{active mean}}$$

$$\sigma^2 = \left(1 + 2\, \underbrace{\frac{k_2}{k_{-1}}}_{\text{burst size}} \left(1 - \underbrace{\frac{k_1}{k_1 + k_{-1}}}_{\text{duty cycle}} \right)^2 \right) \mu^2$$

which have been found previously using different techniques (Peccoud and Ycart, 1995). Here, the burst size is the average number of RNAs produced per an on-off cycle, while the duty cycle is fraction of time spent in the on-state. These expressions make it apparent that the distribution is of high noise, as $\sigma^2 \,/\, \mu^2 \geq 1$ regardless of the parameter values.

S1.2.2 Transcription intervals for the general model. Next, we extend the analysis to the following model:

$$S_0 \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} S_1 \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} \cdots \underset{k_{-n}}{\overset{k_n}{\rightleftarrows}} S_n \xrightarrow{k_{n+1}} E$$

where $n \geq 1$, and the on-off model is recovered as $n = 1$. First, we find the MGFs of the intervals of the state transitions from S_1 to E and S_n to E in terms of a lower order model. For $n = 1$, the two state transitions are equivalent, and the MGF is provided in the previous section. By applying techniques similar to that of the previous section, we get the double recursion:

$${M_n}^{(S_n \to E)}(t) = \frac{1 - p_n}{1 - p_n \frac{v_n}{v_n - t}\, {M_{n-1}}^{(S_{n-1} \to E)}(t)} \frac{v_n}{v_n - t}$$

$${M_n}^{(S_1 \to E)}(t) = {M_{n-1}}^{(S_1 \to E)}\, {M_n}^{(S_n \to E)}$$

where ${M_n}^{(S_1 \to E)}(t)$ and ${M_n}^{(S_n \to E)}(t)$ represent the MGFs of the intervals for the state transitions from S_1 to E and S_n to E, respectively, in a model of order n. Here, $v_n \doteq k_{-n} + k_{n+1}$, which extends the previous definition of v_1, is the exit rate from state S_i. The solutions to the recursion are:

$${M_n}^{(S_n \to E)}(t) = \frac{Q_{n-1}(t)}{Q_{n-1}(0)} \frac{Q_n(0)}{Q_n(t)}$$

$${M_n}^{(S_1 \to E)}(t) = \frac{k_1 - t}{k_1} \frac{Q_n(0)}{Q_n(t)}$$

with $Q_n(t)$ being the continuant:

$$\begin{aligned} Q_0(t) &= k_1 - t \\ Q_1(t) &= (v_1 - t)\,(k_1 - t) - k_1\,k_{-1} \\ Q_n(t) &= (v_n - t)\,Q_{n-1}(t) - k_n\,k_{-n}\,Q_{n-2}(t) \end{aligned}$$

whose roots are real, positive, and distinct, provided that all $k_i, k_{-i} > 0$. Also, in such condition, the roots of $Q_n(t)$ interlace those of $Q_{n-1}(t)$.

Let $p_1 < \cdots < p_{n+1}$ be the roots of $Q_n(t)$. The PDF can be written as:

$$f_n(x) = \sum_{j=1}^{n+1} \left(\frac{k_1 - p_j}{k_1} \underbrace{\prod_{\substack{i \in \{1,\cdots,n+1\} \\ i \neq j}} \frac{p_i}{p_i - p_j}}_{\ell_j(0)} \right) p_j \, \exp(-p_j \, x)$$

where $\ell_j(x)$ is the Lagrange basis polynomial associated with point p_j for the points $p_1, \cdots, p_{n+1}$. In the special case, where k_1 happens to equal one of the poles p_j, this corresponds to the PDF of a sequence of n elementary reactions. Otherwise, a model with equivalent distribution on the state transition duration from S_1 to E would be:

$$S_0 \underset{\tilde{k}_{-1}}{\overset{k_1}{\rightleftharpoons}} S_1 \xrightarrow{\tilde{k}_2} S_2 \xrightarrow{p_3} \cdots \xrightarrow{p_n} E$$

where $\tilde{k}_2 = p_1 \, p_{n+1} \, / \, k_1$ and $\tilde{k}_{-1} = (k_1 - p_1) \, (p_{n+1} - k_1) \, / \, k_1$. This can be found by equating the MGFs of the two models, and as $\tilde{k}_{-1}, \tilde{k}_2 > 0$, this always results in a valid model. It might be possible to select different poles to find $\tilde{k}_{-1}, \tilde{k}_2$, but the above choice is guaranteed to work always, since $p_1 < k_1 < p_{n+1}$. Consequently, for the higher-order models $n > 1$, there are multiple (however, only countably many) model parametrizations of a single interval distribution. Meanwhile, note that it is not possible to identify the original parameters $k_1, \cdots, k_{n+1}, k_{-1}, \cdots, k_{-n}$ as their parameter space is larger than that of $k_1, \tilde{k}_{-1}, \tilde{k}_2, p_3, \cdots, p_n$.

Finally, the MGF for the state transition from S_0 to E is:

$$M(t) = \frac{Q_n(0)}{Q_n(t)} = \frac{p_1}{p_1 - t} \cdots \frac{p_{n+1}}{p_{n+1} - t}$$

which is equivalent to a sequence of $n+1$ elementary reactions with rates $p_1, \cdots, p_{n+1}$ (with no on-off mechanism). Unlike in the case where the rates k_{-i} can be zero, singularities are of no concern as the poles p_i of the MGF are distinct.

S1.2.3 Maximum-likelihood estimation for exponential distribution. Here, we show how to obtain the ML estimator in the presence of both interval- and right-censored exponentially distributed data. This is useful for fitting a model of a single exponential or for acquiring an initial estimate for the mean duration of the other models.

An exponential distribution with a mean of λ^{-1} has a CDF of:

$$F(x) = 1 - \exp(-\lambda \, x)$$

and consequently:

$$\begin{aligned} \ell(\lambda \,|\, x, y) &\doteq \log(F(y) - F(x)) \\ &= \log(1 - \exp(-\lambda \, (y - x))) - \lambda \, x \end{aligned}$$

and $\ell(\lambda \,|\, x, x) \propto \log \lambda - \lambda \, x$ can be found via the limit $F'(x) = f(x)$. Now, ℓ_λ, the partial derivative of ℓ with respect to λ is:

$$\ell_\lambda = \frac{1}{\lambda} \, \frac{\lambda \, (y - x)}{\exp(\lambda \, (y - x)) - 1} - x$$

where the singularities can be removed, resulting in:

$$\ell_\lambda = \begin{cases} \frac{1}{\lambda} - x & \text{, for } y \to x \\ -x & \text{, for } y \to \infty \end{cases}$$

and the second derivative:

$$\ell_{\lambda\lambda} = -\frac{\exp(\lambda \, (y - x)) \; (y - x)^2}{(\exp(\lambda \, (y - x)) - 1)^2}$$

which is negative for finite λ and zero for $\lambda \to \infty$, implying that ℓ_λ is monotonically decreasing and ℓ is concave.

Consider the following Taylor series expansion around $x = 0$:

$$\frac{x}{\exp(x) - 1} = 1 - \frac{1}{2} \, x + \frac{1}{12} \, x^2 + \mathcal{O}(x^4)$$

whose truncation $1 - \frac{x}{2}$ is a lower bound in $x \geq 0$, from which a lower bound for ℓ_λ can be derived. Combining this with the limits of ℓ_λ from above gives:

$$\tilde{\ell}_\lambda = \begin{cases} \frac{1}{\lambda} - \frac{y - x}{2} - x & \text{, otherwise} \\ -x & \text{, for } y \to \infty \end{cases}$$

The estimator using the approximate $\tilde{\ell}_\lambda$ for multiple N iid values (x_i, y_i), where y_i are finite for $i \in \{1, \ldots, M\}$ and right-censored otherwise, is:

$$\frac{1}{\tilde{\lambda}} = \frac{1}{M} \left(\sum_{i=1}^{N} x_i + \sum_{i=1}^{M} \frac{y_i - x_i}{2} \right)$$

which is exact if (x_i, y_i) are not interval-censored.

Now, it must be that $\ell_\lambda \geq \tilde{\ell}_\lambda$ is non-negative at $\tilde{\lambda}$, and $\lambda \to \infty$: $\ell_\lambda \to -x$ is non-positive. As a consequence, there exists a zero of ℓ_λ at $\hat{\lambda}$ such that $0 \leq \frac{1}{\hat{\lambda}} \leq \frac{1}{\tilde{\lambda}}$, which is the ML estimate. This zero can be found using bisection.

REFERENCES

Golding, I. and Cox, E. C. (2004). RNA dynamics in live *Escherichia coli* cells. *Proc. Natl. Acad. Sci. U.S.A.*, **101**(31), 11310–11315.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**(6), 1025–1036.

Kandhavelu, M., Hakkinen, A., Yli-Harja, O., and Ribero, A. S. (2012). Single-molecule dynamics of transcription of the lar promoter. *Phys. Biol.*, **9**(2), 026004.

Le, T. T., Harlepp, S., Guet, C. C., Dittmar, K., Emonet, T., Pan, T., and Cluzel, P. (2005). Real-time RNA profiling within a single bacterium. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(26), 9160–9164.

Lutz, R. and Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucl. Acids Res.*, **25**(6), 1203–1210.

Makela, J., Kandhavelu, M., Oliveira, S. M. D., Chandraseelan, J. G., Jason, L.-P., Peltonen, J., Yli-Harja, O., and Ribeiro, A. S. (2013). *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucl. Acids Res.*, **41**(13), 6544–6552.

Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., Kudasov, F., Chowdhury, S., Yli-Harja, O., and Ribeiro, A. S. (2012). Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucl. Acids Res.*, **40**(17), 8472–8483.

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.*, **48**(2), 222–234.

Svenningsen, S. L., Costantino, N., Court, D. L., and Adhya, S. (2005). On the role of Cro in λ prophage induction. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(12), 4465–4469.

Table S1. Statistics of the models used for the Monte Carlo simulations.

Model	exp	seq-2	seq-3	onoff	onoff-2
Parameters	2750	712, 716	109, 254, 254	360, 1020, 102	360, 1020, 51, 51
Mean (sd)	2750 (2750)	1428 (1010)	617 (375)	138 (212)	120 (143)
Cv-squared	1	0.5	0.37	2.36	1.41
Entropy	8.92	8.15	7.20	5.86	5.67
Cells	60	40	113	100	100
Sampling	19×180	121×60	61×60	241×20	241×20

The table shows the model parameters and the mean, standard deviation (sd), the squared coefficient of variation (cv-squared), and the differential entropy of the resulting intervals. Also shown is the number of cells and the time series sampling settings (samples $\times$ sampling interval), which is used for the time series simulations. Units of time are in seconds, and the entropy is in nats.

Table S2. Performance of model selection and effects of sample sizes in Monte Carlo simulations.

Model	expc	seq-2	seq-3	onoff	onoff-2
Sel (alt) model	exp (seq-2)	seq-2 (onoff)	seq-3 (seq-2)	onoff (exp)	seq-2 (onoff-2)
KLD mean (sd)	0.00507 (0.00679)	0.00624 (0.00706)	0.0085 (0.00863)	0.0223 (0.0274)	0.0696 (0.0217)
LL mean (sd)	-892 (10.1)	-815 (7.88)	-720 (7.88)	-583 (13.4)	-571 (13.8)
Parameter med	2750	688, 744	154, 206, 258	312, 779, 96.8	22.1, 97.4
Choice	0.985	0.943	0.485	0.534	0.512
Sel (alt) model	exp (seq-2)	seq-2 (seq-3)	seq-3 (onoff)	onoff (onoff-2)	onoff-2 (onoff)
KLD mean (sd)	0.001 (0.0015)	0.00143 (0.00177)	0.002 (0.00192)	0.00317 (0.00275)	0.00369 (0.00495)
LL mean (sd)	-4460 (22.4)	-4070 (18.3)	-3600 (17.8)	-2930 (29.7)	-2830 (23.9)
Est med	2760	693, 737	128, 221, 269	355, 982, 101	368, 1050, 50.5, 51.6
Choice	0.973	0.963	0.988	0.836	0.946
Sel (alt) model	exp (seq-2)	seq-2 (seq-3)	seq-3 (onoff)	onoff (onoff-2)	onoff-2 (onoff)
KLD mean (sd)	0.000546 (0.000722)	0.000734 (0.000891)	0.00107 (0.000955)	0.00163 (0.00145)	0.0018 (0.00155)
LL mean (sd)	-8920 (33.1)	-8150 (26.3)	-7200 (25.6)	-5860 (42.7)	-5670 (34.4)
Parameter med	2750	696, 731	125, 222, 270	360, 1020, 102	369, 1040, 50.2, 51.9
Choice	0.981	0.966	0.997	0.83	0.972

Blocks from top to bottom: 100, 500, and 1000 samples. The table shows the most frequently (sel model) and the second most frequently selected model (alt model), the mean (sd) Kullback-Leibler divergence (KLD), the mean (sd) log-likelihood (LL), the spatial median of the parameter estimates, and the frequency of choice for the most frequently selected model. Units of time are in seconds, and the entropy-based measures are in nats.

Table S3. Performance when using different modes of censoring in Monte Carlo simulations.

Model	exp	seq-2	seq-3	onoff	onoff-2
Samples mean (sd)	70.3 (8.31)	191 (9.99)	623 (15.7)	3540 (88.1)	4020 (73.3)
KLD mean (sd)	0.0183 (0.0269)	0.00461 (0.00564)	0.00197 (0.00171)	0.000512 (0.000393)	0.000435 (0.000356)
LL mean (sd)	−87.6 (14.7)	−515 (26.7)	−1260 (28.3)	−7640 (132)	−7860 (90.9)
Parameter med	2670	682, 749	141, 203, 269	357, 989, 101	363, 1040, 51.2, 50.7
Samples mean (sd)	70.3 (8.31)	191 (9.99)	623 (15.7)	3540 (88.1)	4020 (73.3)
KLD mean (sd)	0.0190 (0.0273)	0.00485 (0.00595)	0.00221 (0.00215)	0.000560 (0.000434)	0.000479 (0.000368)
LL mean (sd)	−258 (51.0)	−1240 (74.1)	−3710 (101)	−20200 (448)	−22199.17 (353.51)
Parameter med	2940	677, 766	117, 233, 271	377, 1100, 104	367, 1050, 51.5, 51.0
Samples mean (sd)	28.8 (6.17)	151 (9.98)	510 (15.7)	3440 (88.1)	3920 (73.3)
KLD mean (sd)	1.00 (0.311)	0.0227 (0.0160)	0.0123 (0.00665)	0.00169 (0.000949)	0.000990 (0.000724)
LL mean (sd)	−56.2 (11.7)	−485 (27.6)	−1190 (29.6)	−7500 (135)	−7760 (93.3)
Parameter med	884	617, 629	141, 199, 226	324, 1060, 98.4	330, 1120, 50.7, 50.0
Samples mean (sd)	28.8 (6.17)	151 (9.98)	510 (15.7)	3440 (88.1)	3920 (73.3)
KLD mean (sd)	0.970 (0.295)	0.0227 (0.0158)	0.0122 (0.00652)	0.00168 (0.00535)	0.000960 (0.000709)
LL mean (sd)	−224 (47.4)	−1210 (74.8)	−3630 (102)	−20000 (446)	−22100 (355)
Parameter med	896	617, 629	126, 212, 228	341, 1200, 101	332, 1130, 50.6, 50.5

Blocks from top to bottom: full censoring, no interval-censoring, no right-censoring, and no censoring. The table shows the mean (sd) number of samples per time series, the mean (sd) Kullback-Leibler divergence (KLD), the mean (sd) log-likelihood (LL), and the spatial median of the parameter estimates. Units of time are in seconds, and the entropy-based measures are in nats.

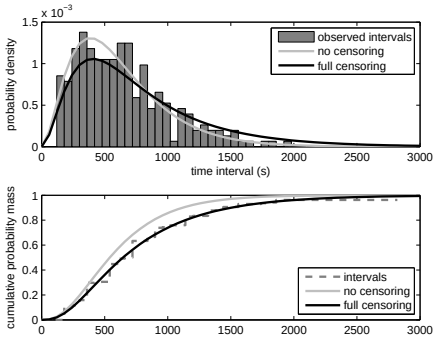

Fig. S1. Transcription intervals for the tetA promoter. Upper panel: the histogram of observed intervals (bars) and the PDFs for no censoring (light gray) and full censoring (black). Lower panel: the Turnbull's CDF estimate of the data (dashed gray) and the CDFs for no censoring (light gray) and full censoring (black).

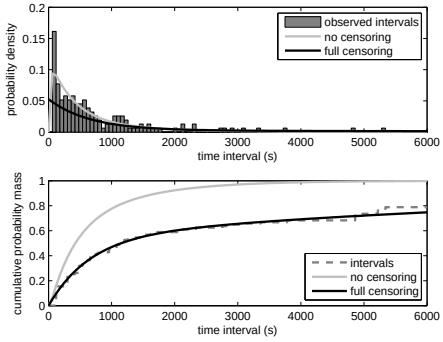

Fig. S2. Transcription intervals for the bacteriophage λ RM promoter. Upper panel: the histogram of observed intevals (bars) and the PDFs for no censoring (light gray) and full censoring (black). Lower panel: the Turnbull's CDF estimate of the data (dashed gray) and the CDFs for no censoring (light gray) and full censoring (black).

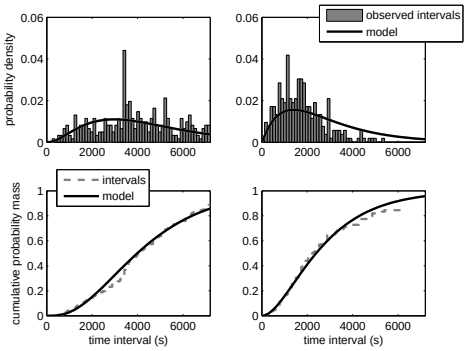

Fig. S3. First RNA production times (left) and transcription intervals (right) for the arabinose BAD promoter. Upper panels: the histograms of the observed data (bars) and the model PDFs. Lower panels: the Turnbull's CDF estimates (dashed gray) and the model CDFs.

Publication IV

Hakkinen, A., Tran, H., Yli-Harja, O., and Ribeiro, A. S., "Effects of rate-limiting steps in transcription initiation on genetic filter motifs," *PLoS One*, vol. 8, no. 8, p. e70439, 2013.

OPEN ACCESS Freely available online

PLOS ONE

Effects of Rate-Limiting Steps in Transcription Initiation on Genetic Filter Motifs

Antti Häkkinen[1], Huy Tran[1], Olli Yli-Harja[1,2], Andre S. Ribeiro[1]*

1 Department of Signal Processing, Tampere University of Technology, Tampere, Finland, **2** Institute for Systems Biology, Seattle, Washington, United States of America

Abstract

The behavior of genetic motifs is determined not only by the gene-gene interactions, but also by the expression patterns of the constituent genes. Live single-molecule measurements have provided evidence that transcription initiation is a sequential process, whose kinetics plays a key role in the dynamics of mRNA and protein numbers. The extent to which it affects the behavior of cellular motifs is unknown. Here, we examine how the kinetics of transcription initiation affects the behavior of motifs performing filtering in amplitude and frequency domain. We find that the performance of each filter is degraded as transcript levels are lowered. This effect can be reduced by having a transcription process with more steps. In addition, we show that the kinetics of the stepwise transcription initiation process affects features such as filter cutoffs. These results constitute an assessment of the range of behaviors of genetic motifs as a function of the kinetics of transcription initiation, and thus will aid in tuning of synthetic motifs to attain specific characteristics without affecting their protein products.

Citation: Häkkinen A, Tran H, Yli-Harja O, Ribeiro AS (2013) Effects of Rate-Limiting Steps in Transcription Initiation on Genetic Filter Motifs. PLoS ONE 8(8): e70439. doi:10.1371/journal.pone.0070439

Editor: Christophe Herman, Baylor College of Medicine, United States of America

Received March 13, 2013; **Accepted** June 18, 2013; **Published** August 5, 2013

Funding: This work was supported by the Academy of Finland and the Finnish Funding Agency for Technology and Innovation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: andre.ribeiro@tut.fi

Introduction

Genes function in networks, whose building blocks are motifs of few genes. Several motifs have been identified, which perform a specific function in networks [1]. Examples include genetic switches, which can be used as memory circuits or for digital control of processes; oscillators, which can be used for time-keeping and synchronization; and genetic filters, which can be used for noise filtering and computation via genetic logic [1].

In addition to the gene-gene interactions, the behavior of a motif depends on the expression pattern of each constituent gene. Investigating this dependency is of relevance given recent evidence that both mean level and the cell to cell diversity in RNA and protein numbers vary between genes by several orders of magnitude [2]. For that, we need to use models that account for the nature of gene expression, since genes with low expression levels are abundant in bacteria [2,3]. Such low numbers cause the dynamics of motifs to be poised with correlations and low copy number fluctuations.

Much effort has been made to characterize the processes of transcription and translation in bacteria. In vitro studies [4,5] showed that transcription, the process by which RNA molecules are produced, is controlled mostly at the promoter region of the gene. Once the RNA polymerase reaches the transcription start site and forms the closed complex, it remains there until the open complex is complete. Following this, the polymerase can escape the promoter and elongate along the DNA sequence, according to which the RNA sequence will be assembled. Both in vitro and in vivo studies suggest that the closed and open complex formations are the lengthiest (rate-limiting) steps of the process of gene expression, along with protein folding and activation.

Recently, the intervals between transcription events in individual, live cells have been measured for two promoters, lac-ara-1 [6] and tetA [7]. These studies suggest that, under optimal conditions, there are two to three major rate-limiting steps, which occur during initiation, that control both mean rate and noise in RNA production. These steps durations were also shown to vary widely with induction level and environmental conditions [6,7]. In that sense, they are major regulators of the dynamics of mRNA production.

Since the duration of the rate-limiting steps in transcription is both sequence-dependent and regulated by activator and repressor molecules, these steps are both evolvable and adaptive to the environment [6]. Since in prokaryotes translation is coupled with transcription, these steps are likely also key regulators of protein numbers [8]. However, it remains unknown to what extent one can tune the behavior of genetic motifs by selecting specific kinetics of initiation of the constituent genes.

In this work, we study the behavior of stochastic genetic motifs, while varying the kinetics of transcription initiation of the constituent genes. Two motifs are considered: one performs filtering in the amplitude domain, and the other in the frequency domain. The response of the motifs is quantified for a wide range of transcriptional dynamics that are in accordance with measurements.

The results indicate that the dynamics of these two genetic motifs, while dependent of the gene-to-gene interactions, is also affected by the kinetics of transcription initiation of each component gene. This, in turn, suggests that it is possible to engineer synthetic circuits to be more robust or having higher plasticity than the present ones, by selecting for promoters with appropriate initiation kinetics.

Methods

Gene expression

We use the delayed stochastic modeling strategy [9,10], which correctly accounts for the low copy number effects, that is, the fluctuations and correlations, of the interacting components, coupled with non-exponential waiting times. The results are quantified from Monte Carlo simulations of the reaction system, using SGN Sim [11].

To model gene expression we use the following set of reactions. The syntax $A \xrightarrow{k} B + C(\tau)$ denotes a reaction where A is transformed into B and C, with a stochastic rate of k. While B is released in the vessel of reactions instantaneously once the reaction occurs, C is released after a delay of τ [10].

$$S \xrightarrow{\infty} S(\tau) + M(\tau) \tag{1}$$

$$M \xrightarrow{d_M} \varnothing \tag{2}$$

$$\varnothing \xrightarrow{k_P M} P \tag{3}$$

$$P \xrightarrow{d_P} \varnothing \tag{4}$$

where $S=1$ ($S=0$) denotes that the promoter is free (occupied), M is the messenger RNA, and P is the protein. Reaction 1 models transcription, Reaction 2 mRNA degradation (d_M being the mRNA degradation rate), Reaction 3 translation (k_P representing the per-mRNA translation rate), and Reaction 4 protein degradation (d_P denoting the protein degradation rate).

The infinite rate set for Reaction 1 derives from the assumption that there is an inexhaustible pool of polymerases (which is a common assumption for bacteria in optimal growth conditions). The delay τ represents the effects of all rate-limiting steps, including the initiation of transcription up to the production of an mRNA. As mentioned, recent evidence suggests that, in *E. coli* under optimal growth conditions, τ is determined to a great extent by the sum of two to three rate-limiting steps, each following an exponential distribution in duration [6,7]. We use $\tau \sim \Gamma(\alpha, \alpha^{-1}\lambda^{-1})$, which denotes that the delay τ is drawn from gamma distribution with a shape of α and a mean of λ^{-1}. Integer values of α indicate that transcription consists of α sequential steps, each with a rate of $\alpha\lambda$. The gamma distribution has a coefficient of variation (the standard deviation over the mean) of $\alpha^{-1/2}$ regardless of the mean (cf. unity of the exponential distribution, which is a gamma distribution with $\alpha=1$). Consequently, values of $\alpha=1$ will result in a Poisson distributed $M \sim \mathrm{Poi}(\lambda d_M{}^{-1})$, while values of $\alpha<1$ result in a more noisy (super-Poisson), and values of $\alpha>1$ less noisy (sub-Poisson) mRNA number dynamics. We note that even if transcription initiation consists of sequential steps of unequal duration, the gamma distribution is still a good approximation. If the steps are of the same order of magnitude, they can be considered approximately equal, else, fast steps can be neglected.

Finally, we let $\lambda \doteq k_M f(X_1,\ldots,X_n)$, where k_M indicates the maximal expression rate of the promoter, and $f(X_1,\ldots,X_n)$: $\mathbb{N}_0{}^n \mapsto [0,1]$ is a regulatory function of the promoter, which depends on substances X_1 through X_n. It is generally not known which steps are affected by which transcription factors, so we assume that each step is affected in an equivalent manner. The choice of these functions is discussed in the next section. Moreover, we let $\mu \doteq k_M d_M{}^{-1} k_P d_P{}^{-1}$, which coincides with the expected protein level of a gene under full expression.

Unless otherwise stated, we use the parameters $k_M d_M{}^{-1}=5$, $d_M=(5\ min)^{-1}$, $k_P d_P{}^{-1}=100$, and $d_P=(60\ min)^{-1}$. These values were selected in accordance with measurements in live *E. coli* [2]. In the results presented, each simulation is ran for 10^6 *min*, and the system is sampled uniformly every 1 *min*. To assess the kinetics of initiation within a realistic range of parameter values, we set the number of rate-limiting steps $\alpha \in \{1,2,3,5,10\}$. The first three have been observed in measurements of mRNA production kinetics in live *E. coli* cells [6,7]. In vitro studies of the kinetics of this process (see e.g. [12]) provide evidence for the existence of, at least, five rate-limiting steps, namely, closed complex formation, three isomerization steps, and promoter clearance. We also study the effects of setting α to 10 to observe the behavior of the model in limit conditions and due to the fact that some of the steps might be non-exponential in duration, thus requiring multiple exponentially distributed steps to be well described.

Gene regulation

The genes are coupled by interactions between their promoter regions and the proteins they express. The activation/repression of a gene is achieved by the binding of the protein expressed by another gene. Once bound, this protein can either degrade while bound, or unbind. While bound, the propensity for the gene to express differs from the unbound case. The activation/repression of gene B by gene A could be represented by the following set of reactions:

$$P_A + T_B \xrightarrow{k_{AB}} T'_B \tag{5}$$

$$T'_B \xrightarrow{k_{AB} K_{AB}} P_A + T_B \tag{6}$$

$$T'_B \xrightarrow{d_P} T_B \tag{7}$$

where P_A denotes the protein product of gene A, $T_B=1$ denotes that the binding site of the gene B for that protein is free, and $T'_B=1$ (implying $T_B=0$) that the binding site is occupied. Here, Reaction 5 models the binding of the activator/repressor molecule P_A to the promoter region of gene B, Reaction 6 its unbinding, and Reaction 7 the degradation of a bound protein. The rate of binding is denoted by k_{AB} and the disassociation constant by K_{AB}.

To simplify the model, we take the limit $k_{AB} \to \infty$. In this limit, the binding of the regulatory proteins is assumed to be much faster than the rate of transcription. It can be found that in this limit, the expectation $\mathbf{E}[T'_B] = (1 + K_{AB} P_A{}^{-1})^{-1}$ if P_A is constant. Following this, to implement the regulation, we vary the transcription rate such that:

$$f_{AB}(A) = \left(1 + (K_{AB} P_A^{-1})^{+d}\right)^{-1} \quad \textit{iff gene A activates gene B} \tag{8}$$

$$f_{AB}(A) = \left(1 + (K_{AB} P_A^{-1})^{-d}\right)^{-1} \quad \textit{iff gene A represses gene B} \tag{9}$$

and

$$f_{ABC}(A,B)=f_{AC}(A)f_{BC}(B) \quad \textit{iff genes A and B regulate gene C} \tag{10}$$

where d denotes the Hill coefficient, which represents the cooperativity of binding, (e.g. $d=2$ can be taken that there are two binding sites for a same type of protein) determining how steep the transition between on- and off-states (e.g. $\mathbf{E}[T'_B]=0$ and $\mathbf{E}[T'_B]=1$) is. Also, the role of the disassociation constant in this context is now apparent, namely, it follows that $\mathbf{E}[T'_B]=0.5$ iff $K_{AB}=P_A$. In our simulations, we use $d=2$, since many proteins are known to function in a dimeric form [13].

Results

Amplitude filtering

We start by examining how the properties of a genetic motif performing amplitude filtering are affected by the transcriptional dynamics. A genetic motif capable of behaving as a biphasic amplitude filter should allow the output to be active only for a certain range of input levels, which allows a process to be trigged by a narrow range of molecular concentration [1]. The region of inputs where the output is active is called the passband and the non-active regions are referred by stopbands. We model a biphasic amplitude filter consisting of four genes as follows. Gene A activates the expression of genes B and C, and gene B activates the expression of gene D, while gene C represses gene D. We model explicitly the expression of genes B through D, while the relative expression level of gene A acts as an input parameter. This is illustrated in Figure 1. Such a circuit was used to explain the narrow range of induction triggering the expression of Xbra in *Xenopus laevis* [14].

We simulate the model for various values of shape α' and rate k'_M of transcription of genes B and C, while the output gene shape and rate are kept constant ($\alpha=2$, $k_M d_M{}^{-1}=5$). This is due to the fact that the effects of changes in α and k_M in the protein distribution of the output gene are more apparent and not related to the internal behavior of the filter, and because it allows the different cases to be easily compared. We set $K_{BD}=0.25\mu'$ and $K_{CD}=0.1\mu'$, which is expected to produce a biphasic response (see Equations 11 through 13). In this, $\mu'=k'_M d_M{}^{-1} k_P d_P{}^{-1}$ denotes the expression rate of genes B and C under full expression. To vary the mean input level, we vary the quantity $\rho=K_{AB}{}^{-1}$ $P_A=10K_{AC}{}^{-1}P_A\propto P_A$.

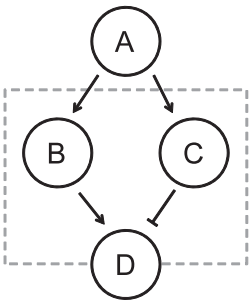

Figure 1. Illustration of the biphasic amplitude filter motif. In the biphasic amplitude filter, gene A acts as input to the filter, while genes C and D compose the filter, represented by the dashed box, along with the regulatory connections between each gene. The protein level of gene D acts as the output.
doi:10.1371/journal.pone.0070439.g001

If all molecule numbers were constant, the response of the filter could be characterized by the following equations:

$$P_B=\mu'\left(1+(K_{AB}P_A{}^{-1})^d\right)^{-1} \tag{11}$$

$$P_C=\mu'\left(1+(K_{AC}P_A{}^{-1})^d\right)^{-1} \tag{12}$$

$$P_D=\mu\left(1+(K_{BD}P_B{}^{-1})^d\right)^{-1}\left(1+(P_C K_{CD}{}^{-1})^d\right)^{-1} \tag{13}$$

which is a good approximation for high expression levels. Note that in Equation 13, P_D is a function of ρ, but invariant to the parameters α' and k'_M, thus the effects of varying them lie beyond this formula. The response of the filter using Equations 11 through 13 is depicted in Figure 2.

The molecular levels will not be constant in our stochastic model. We quantify the noise in molecular levels using Fano factor (the variance over the mean), which is convenient, since Fano factor of Poisson-distributed molecules equals unity regardless of the mean. Even in the limit $\alpha\to\infty$ the protein levels will remain highly noisy (Fano factor $\mathbf{Fano}[P]\geq 1$), since in this case $P_B,P_C\sim\mathrm{Poi}(\mu')$ and their noise further propagates through the probabilistic expression of gene D to the output protein levels P_D.

Next, we present the response of the biphasic amplitude filter using the stochastic model, and study how much it deviates from the expected response when the shape and rate of transcription are varied. The mean output level of the output gene D is presented in Figure 3. As expected, the response resembles the curves in Figure 2. Lower values of α (which imply higher noise) produce slightly degraded performance in terms of the response of the filter. That is, the maximum output protein level will be lower, and the transition between the on- and off-states will be less steep. In addition, the increased noise makes the passband to shift toward a higher input level, since the distributions resulting from the model tend to have right skew.

We also assessed the response for various mean expression levels μ' of the component genes (Figure 4). The results are qualitatively similar to those in Figure 3. Decreasing α' or k'_M (either leading to higher noise) will degrade the filter performance. Moreover, as the expression rate is lowered the shape of the transcription takes greater role in determining the filter behavior. This implies that for rarely expressed genes, it might be important to have sub-Poissonian transcript dynamics, to compensate the increased low copy number noise.

Adding noise in the processes within the filter must shift downwards the value of the maximum output protein level. Generally, adding noise results in a flatter response, which can be interpreted as a degradation in performance, since the filter aims to selectively turn the output on or off. Furthermore, it is possible that adding noise also shifts the input level for which the maximal output is attained or the locations of the transition bands. The results depend on whether the input distributions and the response function of the filter are symmetric or not.

Finally, we assessed quantitatively the effects on the output of having different values of α', for each expression ratio of the input gene shown in Figure 4. For $\mu'\mu^{-1}=0.01$, increasing α' from 1 to 2, causes the output amplitude in the passband to increase by 10.8%. Increasing α' from 1 to 3 causes the output amplitude to

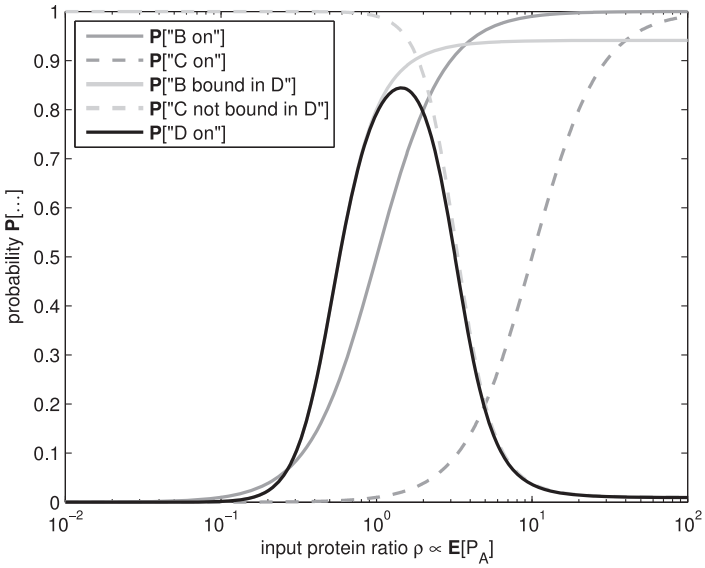

Figure 2. Event probabilities in biphasic amplitude filter. Probabilities of events in the biphasic amplitude filter as a function of the input protein level $\mathbf{E}[P_A]$. The solid black line denotes the probability that the output gene D is expressing, while the dark gray lines denote those of the intermediate genes (solid denoting gene B and dashed gene C). The probabilities that the intermediate genes allow the output gene to express are depicted by the light gray lines (solid denoting gene B and dashed gene C).
doi:10.1371/journal.pone.0070439.g002

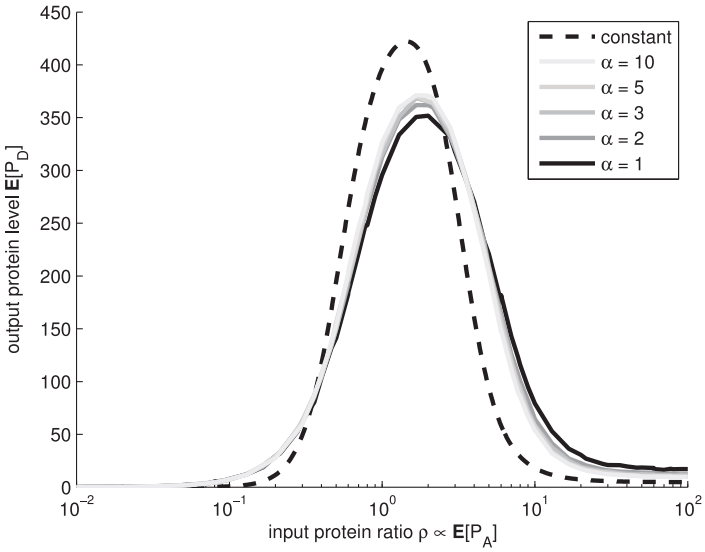

Figure 3. Mean response of biphasic amplitude filter. Mean response $\mathbf{E}[P_D]$ of the biphasic amplitude filter as a function of the protein level $\mathbf{E}[P_A]$ of the input gene, for various shapes α'. Different levels of gray denote different shape parameter α'. The simulations were performed with $\mu'\mu^{-1}$ of 0.05. The dashed black line is an approximation, assuming constant molecular levels.
doi:10.1371/journal.pone.0070439.g003

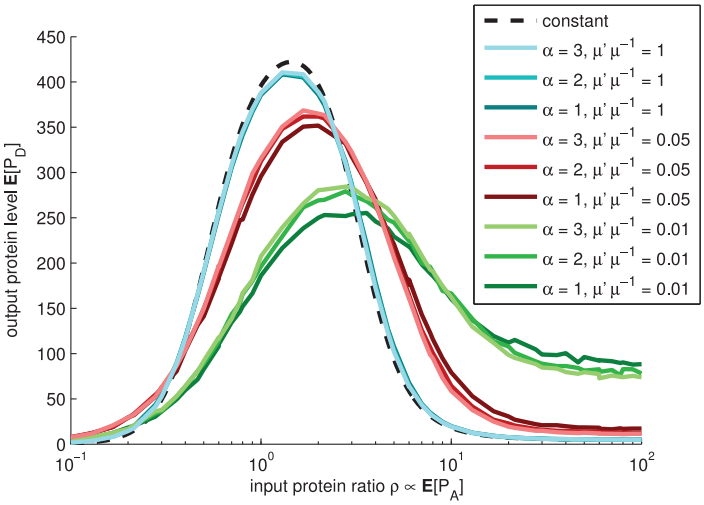

Figure 4. Mean response of biphasic amplitude filter for various transcription rates. Mean response $\mathbf{E}[P_D]$ of the biphasic amplitude filter as a function of the input gene protein level $\mathbf{E}[P_A]$, for various shapes α' and rates k'_M of transcription. Different levels of brightness denote different shape parameter α'. The simulations were performed with $\mu'\mu^{-1}$ of 1 (cyan), 0.05 (red), and 0.01 (green), in the order of decreasing performance. The three cyan lines overlap. We also performed simulations with $\mu'\mu^{-1}$ of 0.5, 0.2, 0.1, and 0.02 (not shown) to assert that the changes are generally nonlinear and more drastic for low mean levels. The dashed black line is an approximation, assuming constant molecular levels.
doi:10.1371/journal.pone.0070439.g004

increase by 12.9%. For other values of $\mu'\mu^{-1}$, the differences are smaller. For example, for $\mu'\mu^{-1}=0.05$, these increases are, respectively, 7.2% and 8.5%, while for $\mu'\mu^{-1}=1$, these differences are of the order of 1.5%.

Since our model dynamics is poised with noise, we study the noise in the output gene protein level, as a function of the input gene level. One might expect the noise to take a shape that is characteristic to the output gene, e.g. constant for Poisson, or some monotonically decreasing curve in our case. In the presence of noisy molecular levels in the circuit, this is generally not true. The noise in the output of this motif is expected to be higher in the transition bands of the biphasic amplitude filter, with the magnitude more characteristic to the output gene in the pass- and stop-bands. An example from stochastic simulations is presented in Figure 5.

From Figure 5 we find that even when the effects of changes in transcription initiation on the response of the biphasic amplitude filter are slight, the change in the fluctuations of the protein numbers of the output gene might be significant. In Figure 6, we present the output noise for various mean levels. For very low expression levels, the low copy number noise in the output becomes dominant.

As a consequence of the amplification of the noise in the transition bands, the output of the filter becomes unpredictable in these regions. Therefore, for this circuit to operate properly in these regions, it is of importance to minimize the noise in the genes composing the filter, for example, by adding rate-limiting steps in initiation. Alternatively, regulation schemes that can provide steeper transition bands are required, which can be accomplished via regulatory schemes of higher-order. We hypothesize that the latter scheme has less effect, since it cannot remove the problem, only reduce its effects. Moreover, it is harder to implement in real genetic circuits, as it requires altering both the protein and the promoter sequences.

Frequency filtering

In this section, we study the effects of changes in the transcription dynamics to a motif that performs filtering in the frequency domain. It is known that changes in the transcriptional dynamics can affect the period and its robustness of genetic oscillators [15], so we expect that these changes affect the response of certain frequency filters as well.

We constructed a motif that can perform low-pass frequency filtering composed of four genes (A through D). This filter suppresses highly transient signals while letting slowly varying signals to pass through as-is. Such a filter would allow a specific set of genes to be subject to only stable signals, by filtering out fast fluctuations in the numbers of the regulatory molecules. Here, gene A acts as an input, required to enable the expression of gene B. Gene B represses gene C, C represses D, and D represses B, that is, genes C through D form a loop (three-gene repressilator). The structure of the motif is illustrated in Figure 7.

When a periodic signal P_A is applied, the behavior of this circuit should vary, depending on the frequency of the signal. When the signal is of high frequency, the feedback loop should be the main responsible for the frequency content of the output. For low frequencies, the input from gene A will disconnect the feedback loop periodically, and lower frequencies, including that of P_A, are introduced in the output. Thus, it is expected that the modulated circuit would have a synchronization point when the input frequency equals that of the repressilator, and that a phase transition would occur in the output frequency response.

For simplicity, we let the Hill coefficient $d'\to\infty$, in the regulatory connection where A activates B. That is, the regulatory connection becomes Boolean, with a threshold of K_{AB}. We denote the Boolean input signal by $X \doteq (1+(K_{AB}P_A{}^{-1})^{d'})^{-1}$. This allows us to omit the explicit modeling of gene A, and consequently this parameter does not need to be determined. Instead, we can apply an arbitrary $X\in\mathbb{B}$. In this case, it does not

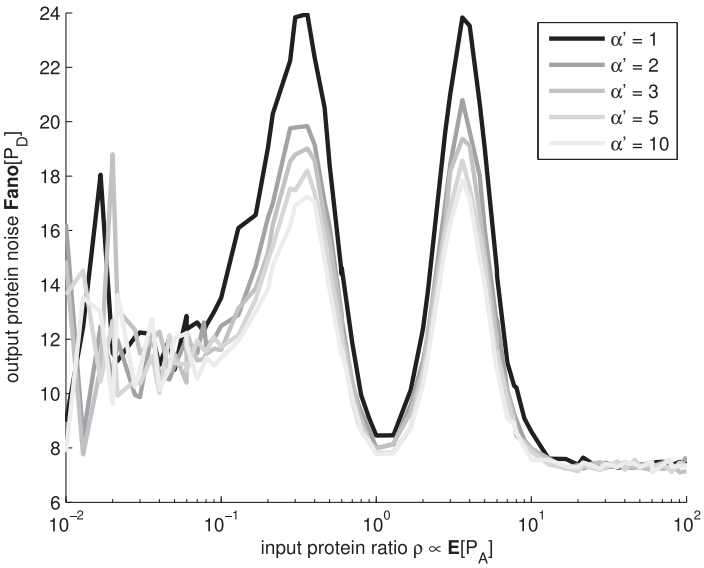

Figure 5. Noise of response of biphasic amplitude filter. Noise of the response $\mathbf{Fano}[P_D]$ of the biphasic amplitude filter as a function of the input gene protein level $\mathbf{E}[P_A]$, for various shapes α'. Different levels of gray denote different shape parameter α'. The simulations were performed with $\mu'\mu^{-1}$ of 1.
doi:10.1371/journal.pone.0070439.g005

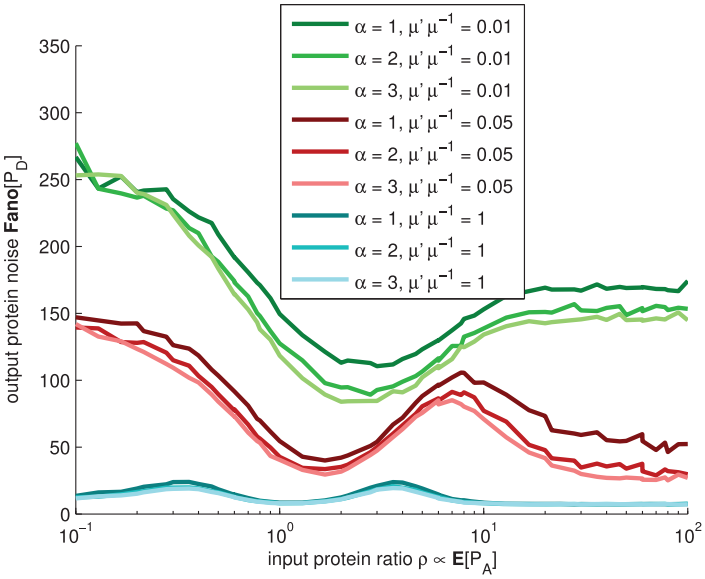

Figure 6. Noise of response of biphasic amplitude filter for various transcription rates. Noise of the response $\mathbf{Fano}[P_D]$ of the biphasic amplitude filter as a function of the input gene protein level $\mathbf{E}[P_A]$, for various shapes α' and rates k'_M of transcription. Different levels of brightness denote different shape parameter α'. The simulations were performed with $\mu'\mu^{-1}$ of 0.01 (green), 0.05 (red), and 1 (cyan), in the order of decreasing noise. We also performed simulations with $\mu'\mu^{-1}$ of 0.5, 0.2, 0.1, and 0.02 (not shown) to assert that the changes are generally nonlinear and more drastic with low mean levels. The dashed black line is an approximation, assuming constant molecular levels.
doi:10.1371/journal.pone.0070439.g006

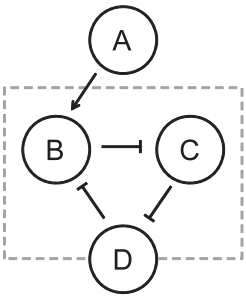

Figure 7. Illustration of the frequency filtering motif. In the frequency filtering motif, gene A acts as an input to the motif, while the filter consists of genes B, C, and D in a feedback loop structure along with the modulation by the input, represented by the dashed box. The protein level of gene D acts as an output of the filter.
doi:10.1371/journal.pone.0070439.g007

matter if the connection is an activating (as in Figure 7) or repressing, since the Boolean input can be flipped.

First, we let the input signal to be constant $X=1$. We analyze the periodic behavior characteristic to the submotif of genes B, C, and D. Since the genes B, C, and D are identical, we can treat them interchangeably and quantify the distribution of periods from any of the protein levels, denoted by T_{BCD}, from the zeros of the autocorrelation function of each time series.

We simulate our model for values of shape α' and rate k'_M of genes B, C, and D, and μ' is defined analogously to the previous subsection. Moreover, the disassociation constants are set to $K_{BC}=K_{CD}=K_{DB}=0.05\mu'$, which were found to produce an oscillatory signal under constant input. The mean period of the protein levels of genes B, C, and D, as a function of the mean expression level μ' of the genes, is shown in Figure 8.

Interestingly, the period changes as a function of the number of steps in transcription initiation. Also, changing the mean transcription level affects the period (note that the disassociation constants are a function of the expected expression level μ', which would make a deterministic model invariant of μ').

We also examined if the robustness of the period is affected. We quantify robustness by the coefficient of variation of the periods of the protein numbers. This measure is convenient, since it equals unity for exponentially distributed periods regardless of the mean. The results are shown in Figure 9. For low mean protein numbers, the period becomes unpredictable (i.e. exponential-like), whereas for moderate levels, the period distribution is Gaussian-like, due to lower noise in transcript production, implying more robust period length. The shapes of the distribution were verified from period histograms (see examples in the insets in Figure 9).

Next, we apply an unbiased Boolean square wave to X, that is, $X(t)=0$ for time t that satisfies $kT\le t<(k+1/2)T$ with any integer k and $X(t)=1$ otherwise, and we denote its frequency by $f_X=T^{-1}$, where T refers to the period. The autocorrelation function of this signal X is a triangular wave of the same frequency, and consequently its spectral power is concentrated to the harmonics of f_X. The spectral power is measured in terms of power spectral density (PSD), which is given by the Fourier transform of the autocorrelation function and measures how much of the signal power per unit frequency is concentrated around certain frequency. Specifically, the PSD of X at frequency f_X is $4\pi^{-2}$ (cf. Figure 10).

We measure the power spectral densities of the input X and the output P_D. An example is shown in Figure 10, with the input PSD plotted for reference. The motif exhibits a low-pass behavior in the frequency domain. Frequencies lower than those corresponding to the mean period of the three-gene submotif when functioning

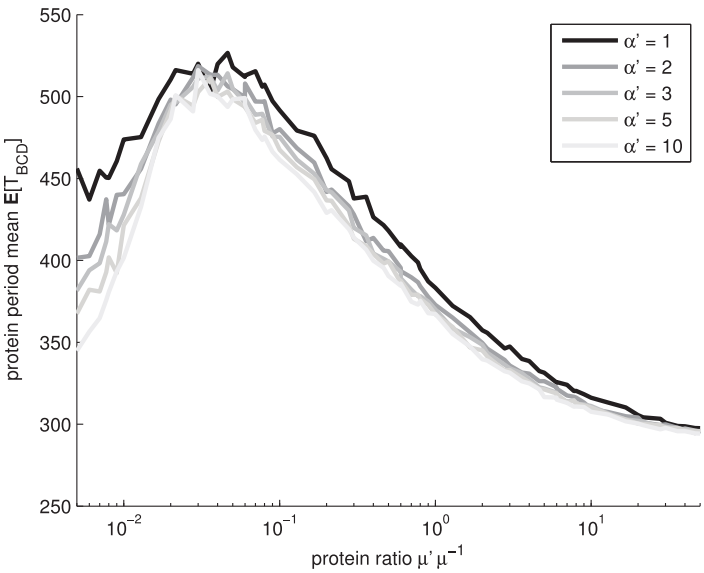

Figure 8. Mean period of frequency filtering motif with constant input. Mean period of the protein levels of genes B, C, and D ($\mathbf{E}[T_{BCD}]$), for constant input $X=1$. Different levels of gray denote different shape parameter α'.
doi:10.1371/journal.pone.0070439.g008

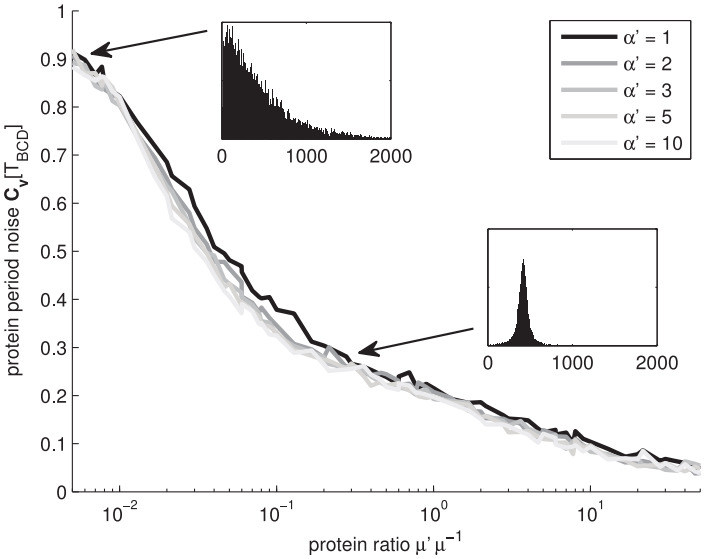

Figure 9. Noise in period of frequency filtering motif with constant input. Noise in the period of the protein levels of genes B, C, and D ($\mathbf{C}_v[T_{BCD}]$), for constant input $X=1$. Different levels of gray denote different shape parameter α'. The insets exemplify the distributions of periods T_{BCD} for shape of $\alpha'=1$ and ratios $\mu'\mu^{-1}$ of 0.005 and 0.5 (units of the x-axis are seconds).
doi:10.1371/journal.pone.0070439.g009

independently (see Figure 8) are only slightly attenuated (amplification factor of $>10^{-1}$). In contrast, higher frequencies are highly attenuated (amplification factor of $<10^{-4}$).

Changing the shape parameter α' of the transcription results in slight variations in the performance of the frequency filter, while the main characteristics are not changed. Namely, the attenuation

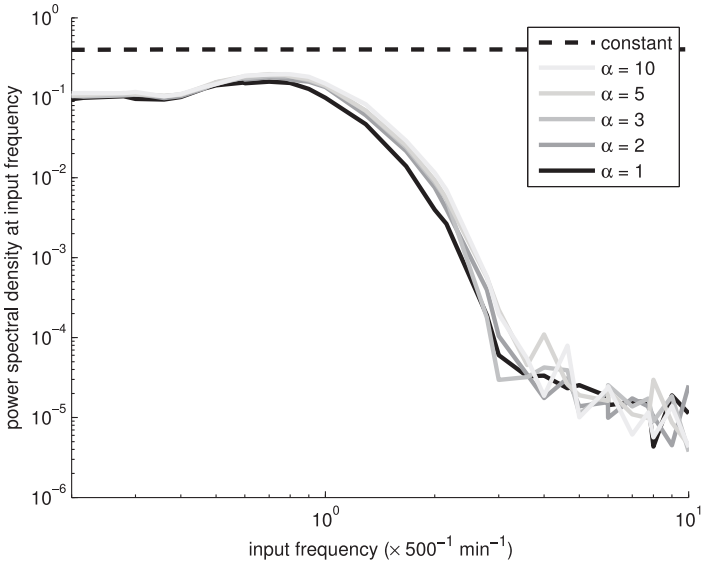

Figure 10. Power spectral density of the frequency filtering motif. Power spectral density of the frequency filter as a function of the input frequency. Different levels of gray denote different shape parameter α'. The simulations were performed with $\mu'\mu^{-1}$ of 0.1. The dashed black line represents the PSD of the input X at the input frequency.
doi:10.1371/journal.pone.0070439.g010

of the frequencies is of the same order of magnitude, more noisy shapes resulting in slightly higher attenuation in the passband. Moreover, the cutoff frequency is affected by changes in the characteristic frequency of the three-gene submotif (Figure 8). We also varied the transcription rate k'_M of the genes in the motif (Figure 11). Again, lower transcription rates, implying more noise in mRNA and protein levels, degrades performance, similarly to when varying α'. The changes in the steepness of the transition band of the filter are more apparent in the former case.

Similarly to the amplitude domain filter, the performance of the frequency domain filter is affected by changes in the transcriptional dynamics of the constituent genes. A transcription process that is less noisy results in a frequency filter with steeper transition bands. Consequently, an efficient frequency domain filter requires limited noise level in transcription, which in the case of low transcript levels can be implemented by a promoter with a sequential initiation process. Interestingly, the cutoff frequency of the filter is also affected by the kinetics of transcription.

As in the case of the amplitude filter, we assessed quantitatively the effects on the output of having different values of α', for each expression ratio of the input gene shown in Figure 11. For $\mu'\mu^{-1}=0.01$, increasing α' from 1 to 2, causes the magnitude of the PSD in the passband to increase by 236.0%. Increasing α' from 1 to 3, causes the PSD to increase by 275.1%. For other values of $\mu'\mu^{-1}$, the differences are smaller as before. In particular, for $\mu'\mu^{-1}=0.05$, these increases are, respectively, 32.5% and 41.9%, while for $\mu'\mu^{-1}=1$, these differences are of the order of 7%.

Discussion

Motivated by recent findings of the relevance of the kinetics of the process of transcription initiation on the dynamics of RNA production in bacteria [6,16], we investigated the functioning of genetic filter motifs as a function of the kinetics of transcription initiation of the constituent genes. We focused on two common filters, namely, an amplitude filter and a frequency filter, as these have several practical applications. One major concern regarding their performance is that most genes in bacteria exhibit very low expression levels. We investigated whether one can utilize the multi-step nature of the process of initiation to compensate for the low copy number noise.

We found that, for realistic parameter values, genetic motifs with stochastic dynamics differ significantly from their deterministic counterparts. Consequently, the latter do not serve as a means to predict the realistic behavior of genetic motifs in live cells. Also, for low expression levels, high noise in the transcripts production significantly degrades the performance of the motifs. The effects of low copy number noise can be compensated by a multi-step (less noisy) transcription process. We suggest that natural motifs with low-expressing constituent genes might employ a multi-step transcription initiation process so as to limit the noise in the mRNA and protein levels, therefore allowing the motif to behave robustly.

The sequence-dependent distribution of transcripts production can have intriguing effects on the behavior of the motifs. These were most prominent in the characteristic frequency of the oscillatory circuit, in which, within a realistic interval of parameter values, it is possible to have a period double that of the one of high

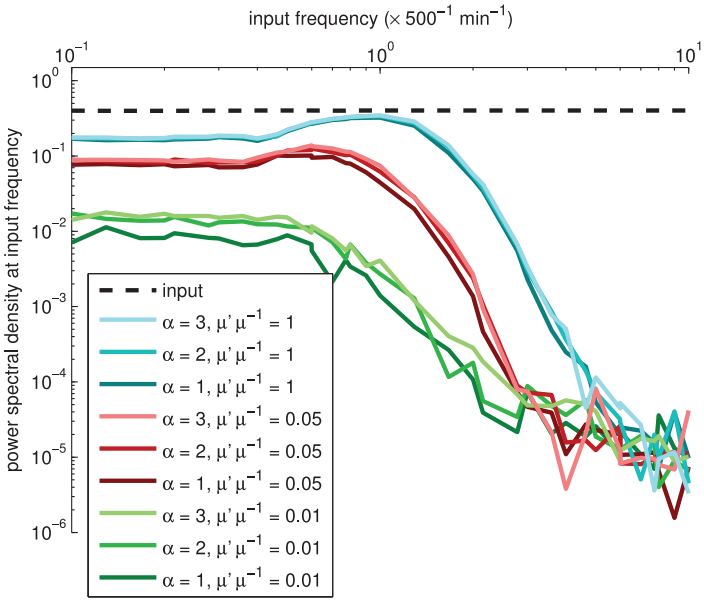

Figure 11. Power spectral density of the frequency filtering motif for various transcription rates. Power spectral density of the frequency filter as a function of the input frequency, for various shapes α' and rates k'_M of transcription. Different levels of brightness denote different shape parameter α'. The simulations were performed with $\mu'\mu^{-1}$ of 1 (cyan), 0.05 (red), and 0.01 (green), in the order of decreasing performance. We also performed simulations with $\mu'\mu^{-1}$ of 0.5, 0.2, 0.1, and 0.02 (not shown) to assert that the changes are generally nonlinear and more drastic with low mean levels. The dashed black line represents the PSD of the input X at the input frequency.
doi:10.1371/journal.pone.0070439.g011

mean levels. Importantly, in both motifs studied, the cutoffs that separate the different regimes of operation of the filters were found to be tunable. The effects of changing the kinetics of transcription initiation were found to be slight, partly masked by the noise, but non-negligible.

It is known that changes in the kinetics of the sequential process of transcription initiation affect the dynamics of mRNA abundances of individual genes [16,17]. Here, we provided tentative evidence that these changes affect the behavior of genetic motifs as well. This is of relevance, since both the number and the kinetics of these steps are dependent of the promoter sequence and transcription factors alone, i.e., are independent of the protein coding region. Due to this, we hypothesize that it is possible to alter the kinetics of a genetic circuit significantly by replacing the promoter region of the constituent genes, without the need of altering the protein under their control. Further, we hypothesize that changes in the promoter sequence of the constituent genes of motifs constitutes a significant degree of freedom in their evolutionary process in natural organisms.

Author Contributions

Conceived and designed the experiments: ASR HT AH. Performed the experiments: ASR HT AH. Analyzed the data: ASR HT AH. Wrote the paper: ASR HT AH OYH.

References

1. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. Curr Opin Microbiol 6: 125–134.
2. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, et al. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329: 533–538.
3. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc Natl Acad Sci USA 99: 9697–9702.
4. McClure WR (1985) Mechanism and control of transcription initiation in prokaryotes. Annu Rev Biochem 54: 171–204.
5. Lutz R, Lozinski T, Ellinger T, Bujard H (2001) Dissecting the functional program of *Escherichia coli* promoters: The combined mode of action of lac repressor and arac activator. Nucl Acids Res 29: 3873–3881.
6. Kandhavelu M, Mannerstrom H, Gupta A, Hakkinen A, Lloyd-Price J, et al. (2011) *In vivo* kinetics of transcription initiation of the lar promoter in *Escherichia coli* : Evidence for a sequential mechanism with two rate-limiting steps. BMC Syst Biol 5: 149.
7. Muthukrishnan AB, Kandhavelu M, Lloyd-Price J, Kudasov F, Chowdhury S, et al. (2012) Dynamics of transcription driven by the *tetA* promoter, one event at a time, in live *Escherichia coli* cells. Nucleic Acids Res 40: 8472–8483.
8. Pedraza JM, Paulsson J (2008) Effects of molecular memory and bursting on fluctuations in gene expression. Science 319: 339–343.
9. Ribeiro AS, Zhu R, Kauffman SA (2006) A general modeling strategy for gene regulatory networks with stochastic dynamics. J Comp Biol 13: 1630–1639.
10. Roussel MR, Zhu R (2006) Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression title. Phys Biol 3: 274–284.
11. Ribeiro AS, Lloyd-Price J (2007) Sgn sim, a stochastic genetic networks simulator. Bioinf 23: 777–779.
12. deHaseth PL, Zupancic ML, Record MT Jr (1998) Rna polymerase-promoter interactions: The comings and goings of rna polymerase. J Bacteriol 180: 3019–3025.
13. Xia K, Manning M, Hesham H, Lin Q, Bystroff C, et al. (2007) Identifying the subproteome of kinetically stable proteins via diagonal 2D SDS/PAGE. Proc Natl Acad Sci USA 104: 17329–17334.
14. Dyson S, Gurdon JB (1998) The interpretation of position in a morphogen gradient as revealed by occupancy of activin receptors. Cell 93: 557–568.
15. Loinger A, Biham O (2007) Stochastic simulations of the repressilator circuit. Phys Rev E 76: 051917.
16. Kandhavelu M, Häkkinen A, Yli-Harja O, Ribeiro AS (2012) Single-molecule dynamics of transcription of the lar promoter. Phys Biol 9: 026004.
17. Ribeiro AS, Hakkinen A, Mannerstrom H, Lloyd-Price J, Yli-Harja O (2010) Effects of the promoter open complex formation on gene expression dynamics. Phys Rev E 81: 011912.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3685-4
ISSN 1459-2045